Development of Network-Analysis Tools and Applications in Biochemistry

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Philipp Schütz

aus

Blumenstein BE

Promotionskomitee

Prof. Dr. Amedeo Caflisch (Vorsitz)
Prof. Dr. Benjamin Schuler

Zürich 2009

Summary

Networks have been widely used in the last decade to analyze large-scale data sets. This popularity originates from the possibility both to represent pairwise interactions among arbitrary objects in simple, two dimensional plots and to treat any type of data with the same formalism. In the first part of this thesis, network technology is applied to examine the free-energy surface of a protein using only a time series of an one-dimensional signal, e.g. an intramolecular distance. In the second part, a new procedure to identify tightly connected communities in large networks is presented. Common to both procedures is the attempt to infer structures in the examined data by analyzing the associated networks.

A complete characterisation of the free-energy surface is essential to understand the folding mechanisms of a protein. With the experimental technique "Förster Resonance Energy Transfer" (FRET) a single intramolecular distance can be monitored with high time resolution over an extended period of time. To assess the information content on the free-energy surface of a FRET experiment, a new method called FESST (Free Energy Surface from Single-molecule Time series) to extract free-energy basins from the time series of a single distance is presented. The central assumption behind FESST is that the distribution of the signal in small time windows is characteristic for the actual free-energy basin. Applied to Beta3s, a 20-residue peptide with native three stranded antiparallel β-sheet conformation, FESST extracts the native state with more than 96% accuracy from a time series of the distance between two C_β-side chain atoms taken from a molecular dynamics simulation. Furthermore, FESST extracts the barrier between folded and unfolded state correctly up to a difference of two percent and detects three additional free-energy basins stabilized mainly enthalpically. Extrapolating the amount of data required by FESST to single molecule FRET experiments, the free-energy basins of proteins with a folding time of few milliseconds can be detected by FESST.

In the second part, a new procedure to detect modules in a large network is presented. To identify groups of nodes with many internal and few inter-community connections, the partition with highest "modularity" has to be identified. This scoring function is used, because it allows an objective and algorithm independent definition of a community. Here, an optimization strategy called "MSG-VM" is presented that is both effective and efficient. For multiple large benchmark networks, MSG-VM improves the literature values for the highest modularity found. Furthermore, a new benchmark network is suggested that represents coappearing words in the title of publications authored by the famous physico-chemist Martin Karplus. Despite the large overlap in vocabulary of the various topics of the work of M. Karplus, the identified groups of words could be attributed to the different fields with ease.

Zusammenfassung

Netzwerke wurden in den vergangenen 10 Jahren vermehrt zur Analyse umfangreicher Datensätze herangezogen. Diese Beliebtheit rührt von der Möglichkeit Paarwechselwirkungen zwischen beliebigen Objekten leicht darzustellen und mit demselben Formalismus zu behandeln. Im ersten Teil der vorliegenden Dissertation werden Netzwerke verwendet, um die Freie Energie Landschaft eines Proteins zu untersuchen unter alleiniger Verwendung der Zeitreihe eines eindimensionalen Signals, beispielweise der Distanz zweier Proteinatome. In einem zweiten Teil wurde ein neues Verfahren entwickelt, um Gruppen mit dichter interner Vernetzung innerhalb von grossen Netzwerken zu identifizieren. Beiden Verfahren extrahieren eine Struktur der zugrundeliegenden Daten aufgrund der Analyse der zugehörigen Netzwerke.

Die vollständige Charakterisierung der Freien Energie Landschaft ist essentiell für das Verständnis der Proteinfaltung. "Förster Resonance Energy Transfer" (FRET) erlaubt die experimentelle Bestimmung einer einzelnen intramolekularen Distanz über einen ausgedehnten Zeitraum mit hoher zeitlicher Auflösung. Um zu untersuchen, wieviel Information über die Freie Energie Landschaft aus FRET Experimenten gewonnen werden kann, wurde eine Methode genannt FESST (Free Energy Surface from Single-molecule Time series) entwickelt, die Basins der Freien Energie Landschaft aus der Zeitreihe einer einzelnen Distanz bestimmt. Die Schlüsselidee für FESST ist, dass die Verteilung der Distanzwerte innerhalb eines kurzen Zeitfensters charakteristisch für das jeweilige Basin der Freien Energie ist. Für das Testsystem Beta3s (synthetisches Peptid mit 20 Amminosäuren und drei antiparallelen β-Faltblättern in der nativen Struktur) extrahiert das vorgestellte Verfahren das native Basin von Beta3s mit mehr als 96 % Genauigkeit. Darüber hinaus sagt die Analyse der Distanzzeitreihe die Barriere zwischen gefaltetem und ungefaltetem Zustand bis auf zwei Prozent genau vorher und identifiziert drei weitere enthalpische Basins. Durch Analyse einer Photonentrajektorie aus einem simulierten FRET Experiment zeigt sich, dass FESST Freie Energie Basins von Proteinen mit einer Faltungszeit von wenigen Millisekunden finden kann.

In einem zweiten Teil wird ein Verfahren zur Detektion von Modulen innerhalb eines Netzwerks diskutiert. Um diese Gruppen von Knoten mit vielen internen und wenigen externen Verbindungen zu finden, wird die Partition mit höchster "Modularität" gesucht. Diese Bewertungsfunktion wird verwendet, um eine objektive und von Algorithmen unabhängige Definition eines Moduls zu erhalten. Der hier diskutierte Optimierungsalgorithmus "MSG-VM" erwies sich im Vergleich zu anderen Algorithmen als effizient und effektiv. Bei vielen gebräuchlichen Testnetzwerken konnten die Literaturwerte sogar übertroffen werden. In einem linguistischen Beispiel wird MSG-VM angewandt, um Wortgruppen in Titeln von Publikationen des berühmten Chemikers Martin Karplus zu identifizieren. Trotz des grossen Überlaps im Vokabular der einzelnen Teilgebiete, konnten die gefundenen Wortgruppen leicht den unterschiedlichen Arbeitsbereichen zugeordnet werden.

List of publications

Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement.
P. Schuetz and A. Caflisch
[*Phys. Rev. E*, **2008**, *77*, 046112]

Multistep greedy algorithm identifies community structure in real-world and computer-generated networks.
P. Schuetz and A. Caflisch
[*Phys. Rev. E*, **2008**, *78*, 026112]

Free energy surfaces from single-distance information
P. Schuetz, B. Schuler, and A. Caflisch
[*submitted*]

Contents

Chapter 1

Applications of networks in the biochemical context

1.1 Outline

In the last decade, networks emerged as a new promising tool to analyze various biochemical problems [1–3]. One major field of application is the analysis of *in silico* protein folding. Monte Carlo and molecular dynamics simulations (section 1.3) model proteins at atomistic level of detail and can elucidate aspects of the fundamental question how proteins fulfill their ubiquitous and diverse functions in an organism (section 1.2). To characterize the folding process, descriptors for its progress and energetics are necessary. The discussion of analysis methods is split into two parts to highlight the impact of network technology: Section 1.4 presents recent approaches without networks, whereas in section 1.5 new methods based on network technology are presented (see [4] for procedures published before 2005).

Förster Resonance Energy Transfer (FRET) is an experimental technique to monitor an intramolecular distance over time. In section 1.6, we review the fundamental idea of a FRET experiment and theoretical methods to infer distance time series from experimental data. Combining network technology and analysis methods for simulations, free-energy basins can be detected from the time series of a single distance.

In living organisms, many processes (e.g. metabolism) are organized in functional modules, i.e. the system is composed of weakly connected subsystems with independent functionality. The last section of this chapter discusses different approaches to detect modules in data sets that can be mapped on networks.

1.2 Proteins: machinery of life

Proteins are unbranched, organic chain molecules assembled from 20 different building blocks called amino acids. These chain molecules participate in almost every process of a living organism. For most proteins, a correct operation is only possible if all amino acids are aligned in a unique three-dimensional arrangement called the native structure. A protein is created at the ribosomal complex by successively appending each amino acid in the sequence at the end of the nascent chain. On its way to the native structure, the protein starts as an unstructured polypeptide chain in the solvent tethered at one end to the ribosomal complex. Levinthal argued that folding cannot be a random reorganization process, because the exploration of the vast space of different protein conformations would exceed the observed folding times [5]. Zwanzig et al. observed that the folding time can be reduced to the biologically relevant range if a weak bias against unfavorable conformations is introduced that originates from intra-protein interactions [6]. The idea of a bias has been extended to the picture of an energy landscape with several funnels corresponding to different (meta-)stable states [7–13]. Key to the resolution of Levinthal's paradox is that the energy landscape governing the folding process is not entirely flat.

Thermodynamically protein folding is described by the Gibbs free energy $G = U - TS$ with U the enthalpic contribution (interaction energy within the protein and between protein and solvent), T the temperature of the system and S the entropy accounting for the flexibility of the protein. The Gibbs free energy is the most suitable thermodynamic potential, because in most experiments and living organisms the number of particles, the ambient temperature, and the environmental pressure stay constant. At equilibrium, the most populated structure minimizes the free energy (visiting probability follows a Boltzmann distribution). At physiological conditions, i.e. temperature, pressure, ionic strength etc. matching those in a living cell, the free energy is minimal for the native state. In non-physiological environments, the most probable structure may completely differ from the native one. For instance, most proteins unfold in a solvent with high concentration of guanidinium chloride. Random collisions with the solvent cause a continuous change and interconversion of entropy and enthalpy. The interconversion rate between two structures is mainly determined by the height of a possible barrier in between [14–16]. The native basin contains all structures that fold to the native state without crossing a barrier. Furthermore, the conformations of the native basin interconvert fast with the native structure (and among them) in equilibrium dynamics. Remarkably, if a barrier has to be crossed, the transition path (sequence of structures visited in the passage) chosen by the

system most likely minimizes the free-energy as well [17, 18]. In conclusion, all relevant processes are governed by the free-energy landscape. Thus, to focus only on the potential-energy landscape is insufficient [19–22].

In the next section, different approaches to study protein folding *in silico* are reviewed.

1.3 Computational methods to study protein folding

The dynamics of a protein can be elucidated by atomistic simulations that require a known three-dimensional protein conformation. If no structure is available, either a folding simulation starting from an extended conformation (linear chain molecule) or a structure prediction procedure can be applied. The popular ROSETTA algorithm tries to predict the native structure of a protein by combining structural motives of the contained amino acid sequences [23–25]. The key assumption is that short sequences fold into a small family of distinct structures. The database linking sequences and distinct folds is constructed from the analysis of all protein structures deposited in the RCSB PDB data base [26].

If a three-dimensional structure of the protein is available or predicted, the conformation space of the molecule can be explored by Monte Carlo (MC) or Molecular Dynamics (MD) simulations. In MC simulations, the protein conformation is iteratively changed according to a prescription chosen at random from a set of transformations called move set. Whether the changed structure is used in the next step is determined according to an acceptance criterion. In general, the Metropolis criterion is applied [27, 28]: If the interaction energy for the protein and/or the solvent decreases upon modification, the transformation is accepted. If the change in interaction energy ΔE is positive, the move is accepted with probability $e^{-\Delta E/k_B T}$ where k_B is Boltzmann's constant and T the temperature of the protein/solvent. MC simulations are still commonly used to minimize the enthalpy of a protein [29–31]. This simulation technique allows fast sampling of the conformation space with the appropriate move set, but does not provide any information on the dynamics. Therefore, MC studies became deprecated for folding studies.

Dynamical information can be gained by MD simulations as discussed in the next section.

1.3.1 Molecular Dynamics simulations

In a molecular dynamics (MD) simulation, each atom of the protein is modelled as a massive and possibly charged point particle in a classical potential. Quantum mechanical effects such as bond lengths, torsional and dihedral angles are treated as individual terms in the potential called force field. The classical trajectories of each atom are obtained solving the equations of motion according to Newton's second law [32]. In commonly used force fields such as CHARMM [33], OPLS/AA [34,35], and Amber [36] harmonic potentials are used to restrain bond length and torsional angles. Changes in the dihedral angles of certain atom types are limited by the inclusion of a washboard potential (energy function with two equally deep wells and one deeper well mimicking one preferred and two alternative angle settings). Electrostatic interactions between charged particles are considered by the classical Coulomb field. The Lennard-Jones potential represents the van-der-Waals force and the repulsion between atoms at short distances. The van-der-Waals forces, i.e. the effect of fluctuating electron clouds, are modelled as the mutual interaction energy of two induced dipoles (energy scales as r^{-6} with r the distance between the two atoms). The r^{-12}-term (r the distance between the two atoms) in the Lennard-Jones potential [37] mimics the repulsion of two atoms at short distance due to the Pauli repulsion of overlapping electron clouds [38]. A particularity of the CHARMM force field is the Urey-Bradley term that fixes the distance between atoms linked to the same central atom by a harmonic potential reducing the effects of angle bending. Although these classical approximations (quadratic potentials etc.) appear very crude, the experimental results can be reproduced reliably [39–43]. Noteworthily, the parameters are optimized to reproduce the biophysical properties of a set of model molecules best for one given temperature and pressure. The differences among the force fields concern the set of molecules and data used for parametrization. The computational demand of the simulation can be reduced, if the number of modelled atoms is lowered. As first approximation, apolar hydrogen atoms can be omitted in proteins, because they cannot form hydrogen bonds as polar hydrogen atoms. For instance, the CHARMM force field is available in three different "flavors": In CHARMM param 19 [44, 45] only the polar hydrogen atoms are modelled explicitly. Both param 22 [46,47] and param 27 [48] model all hydrogen atoms, but param 27 yields more accurate results if DNA, RNA and lipid molecules are included in the simulation.

The folding can be biased towards a particular conformation by a Go model [49], in which only the interactions specific to a target structure (henceforth called contacts) are considered in the potential. A welcome side-effect is that this restriction allows a significant speed-up of the simulation due to

both the biased dynamics and the reduced number of energy terms to calculate. Draw-backs of a Go model are the ambiguity in construction, i.e. several sets of contacts can be used to define the target structure, and the limited exploration of non-target structures.

Each force field is an approximation of the real interaction energy at a fixed temperature and pressure. For other conditions, the interaction energy may be incorrectly reproduced. This implies that the experimental as well as simulation temperature and pressure may differ to reproduce the same behavior. In any case, many examples illustrate that simulations predict the correct change in behavior. For instance, MD simulations of the Trp-cage predict the correct melting temperature [42].

A solution of Newton's equations with the aforementioned force fields has constant energy. In experiments not the energy, but the temperature and the pressure are kept constant. Therefore, additional algorithms such as the Berendsen [50] and Nosé-Hoover [51, 52] thermo- and barostats control the temperature and pressure. Both thermostat algorithms rescale the velocities to fulfill the equipartition theorem [53,54]. For instance, with the Berendsen thermostat the actual simulation temperature converges exponentially to the target temperature with a relaxation time $\tau_{\text{Berendsen}}$. Most barostats control the system pressure by rescaling the spatial coordinates. For further details, the reader is referred to the introductory text [55] as well as to the original publications [50–52].

Often, interactions between protein and solvent change the stability of certain conformations and therefore, are crucial to a correct modelling of the protein dynamics. For instance, in [56] the behavior of water molecules inside the rat intestinal fatty acid binding protein (I-FABP) is studied. The MD simulations predict two long-lived water molecules and on average twenty-two water molecules within the cavity consistent with the NMR measurements. Each water molecule can be explicitly simulated by models such as SPC (three simple point charges with tetrahedral bond angle), and TIP4P [57](additional dummy atom improving the electrostatic distribution). The number of water molecules exceeds the number of protein atoms in general by zero to two orders of magnitude to reduce spurious effects from rotating proteins or mirror artifacts (most proteins are simulated with periodic boundary conditions, i.e. each atom leaving the simulation box is moved to the opposite site).

Water molecules in the simulation increase the number of interactions to calculate significantly. Although the simulations with explicit treatment of water molecules yield a striking agreement with experimental data [39–43], the high computational demand limits the simulation time to few hundreds of nanoseconds. For most systems, the folding time is several orders of mag-

nitude higher, which renders explicit water simulations impractical. Longer simulation times become feasible if the water molecules are eliminated from the simulation and their average effect is mimicked by additional terms in the potential. Effects of corpuscular water such as hydrogen bonds stabilized by an intermediate water molecule cannot be described anymore. But, approximations such as EEF1 [58], SASA [59] and FACTS [60] attain a remarkable accuracy in reproducing biophysical properties of the proteins studied. The recent solvation model FACTS [60] reproduces the crystallographic B-factors for the chymotrypsin inhibitor 2 qualitatively correctly. Apart from reduced computational complexity, the analytical treatment yields a smaller viscosity. Therefore, the protein explores the free-energy landscape much faster. With these modifications it is possible to simulate several folding and unfolding events of a small peptide. As another example, in the SASA [59] model the proportionality of solvation free-energy and solvent accessible surface is exploited. Applied to the peptide Beta3s [61] (Fig. 1.1), a 20-residue protein with native three β-strand topology, the ratio of the formation times for structural elements (helices, β-hairpin, and β-sheets) match the experimental data [62]. Even for the complex system of a small helical peptide with an attached linker (photoswitch to induce folding), the ratio of folding and unfolding times is correctly reproduced [43].

For large proteins, even implicit solvent treatment cannot accelerate the simulations enough. In this case, three different approaches might be used: Transition path identification, adaptive sampling and replica exchange. Transition path identification subsumes several techniques such as transition path sampling [63, 64] and transition path-steered molecular dynamics [18] that identify the most likely transition path between two structures of distinct free-energy basins. The unfolding process of the β-hairpin in the protein G–B1 is examined with transition path sampling in an explicit water simulation [65]. The absolute rate constant for unfolding ($k = 0.2\ \mu\mathrm{s}^{-1}$) of this complex system at 300 K is in reasonable agreement with the experimental data ($k = 0.17\ \mu\mathrm{s}^{-1}$).

If the initial and final structures are unknown, the adaptive sampling techniques are one tool of choice. In flooding [66], accelerated molecular dynamics [67] and stochastic tunneling [68] (predominantly used for MC simulations), the potential is iteratively changed by additional terms that prohibit a further visit of the same point in the conformation space. The residence time in deep basins is then reduced and transition regions are sampled better. Alternatively, the sampling is improved if all potential wells lower than a given energy threshold are filled up [67]. For a hepta-alanine, a standard MD simulation only resolves the basin of the initial helical structure, whereas the accelerated MD scheme (with the same number of MD steps and the same

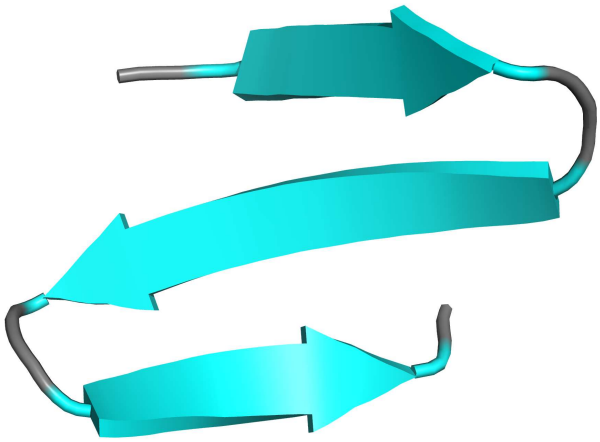

Figure 1.1: Native state conformation of Beta3s. NMR studies [61] of the conformation in solution indicate that Beta3s is a monomeric, three-stranded antiparallel β-sheet peptide. The turns are formed by Gly_6-Ser_7 and Gly_{14}-Ser_{15}.

computation time) is able to resolve up to two additional basins depending on the energy threshold.

Alternatively, the number of transitions can be increased by replica exchange molecular dynamics (REMD) [69]. In REMD simulations, several simulations at different temperatures are run in parallel. After a defined time (usually on the order of 1 - 20 ps), a Metropolis criterion (cf. the beginning of this section) is used to decide whether the structures of the simulations at two neighboring temperatures are exchanged. The central idea behind a REMD simulation is that a barrier prohibiting a transition at lower temperatures can be surmounted at higher temperature. Simulations at low temperatures, on the other hand, are used to explore the basins. Applied on the β-hairpin protein, REMD simulations determine a β-hairpin population of 80% at 270 K, what matches very well the experimental value of 71% determined by fluorescence quantum yield experiments [70]. Although REMD simulations may help to overcome sampling problems, they failed to predict the melting temperature of the Trp-cage more accurately than conventional MD simulations [42]. Folding times cannot be calculated directly in REMD simulations, because each structure is only simulated for a short time (in general only a few multiples of the time between two exchange attempts) at the same temperature. A recent approach [71] combines the dynamical information at different temperatures by a rescaling procedure and is able to reproduce the distribution of folding times for Beta3s.

1.4 Descriptors of protein folding

Folding a protein requires the collaboration of many degrees of freedom. To quantify the evolution of the folding process, descriptors are needed that are discussed in this section.

1.4.1 Order parameters, progress variable and reaction coordinates

An order parameter is "a normalized parameter that indicates the degree of order in the system" [72]. The order parameter value 0 is attributed to total disorder in the system, a value of 1 corresponds to the state of complete order. In protein folding, an order parameter describes the proximity to the native state, which is the "most ordered" state. For Go-models (cf. section 1.3.1), a natural choice for an order parameter is the fraction of native contacts Q. Obviously, the fraction of native contacts is equal to one for all native-like structures. In the denatured state, i.e. the state least similar to the native one, no native contacts are formed and therefore Q vanishes. Structurally spoken, the more native contacts are formed the more native-like the conformation is. If a protein folds by forming contacts successively without breaking existing ones, Q measures the progress of folding. Normalized progress variables (measure for the progress of folding) are called reaction coordinates. In the example above, Q is a reaction coordinate. Although deceivingly simple, Q is not a generic reaction coordinate. For instance, structures in the native state of Beta3s (Fig. 1.1), a 20-residue peptide with native three-strand antiparallel β-sheet conformation, have $Q \sim 0.3$ in a simulation with an implicit solvent treatment in a transferable force field [62]. On the other hand, there exist Beta3s conformations from the denatured state with $Q \sim 0.7$ [62]. Evidence for the inappropriateness of Q as reaction coordinate are supported also by other types of simulations [73]. Key for the existence of a reaction coordinate is that all orthogonal degrees of freedom involved in the folding process relax much faster than the motion along the reaction coordinate [74]. A widely used measure for kinetic distance is the folding probability p_{fold} [73]. To calculate p_{fold} for a given snapshot (taken from a Monte Carlo or molecular dynamics trajectory), a large number of simulations with different initial velocities is started. The folding probability is then calculated as the fraction of simulations that reach the folded state before unfolding. The transition state ensemble - defined as the set of structures that fold or unfold with equal probability - is determined by the set of structures with $p_{\text{fold}} = 0.5$. For the proteins c-src SH3 and CI-2, the folding probability predicts a transition state that is in perfect agreement with the

transition state obtained by other structure based descriptors [75]. Noteworthily, this agreement supports the two-state folder hypothesis for c-src SH3 and CI-2.

Although the folding probability is the least controversial reaction coordinate, the transition state identified for the cyanovirus-N does not match the chemically relevant one [75]. As second draw-back, the evaluation of the folding probability is computationally very demanding, because a significant amount of simulations (up to 100 even for small proteins) is needed for every snapshot. The computational demand can be reduced to the determination of a single trajectory with multiple folding and unfolding events, if an approximation of p_{fold} called cluster-p_{fold} is calculated [76]. For this approximation, the snapshots are structurally clustered and cluster-p_{fold} is given by the fraction of snapshots in the considered cluster proceeding to the folded state before unfolding. In [62], this calculation procedure was successfully applied to determine the Φ-values of Beta3s and 32 single-point mutations thereof by means of folding-unfolding simulations. An analysis of the transition state structures (determined by $p_{\text{fold}} = 0.5$) revealed the presence of specific non-native interactions for most peptides [62].

1.4.2 Constructing reaction coordinates

Although the folding probability measures the progress of the folding accurately, no information can be gained about what structural features are important. Ma and Dinner suggested a procedure to identify the optimal combination of coordinates (interatomic distances and dihedral angles) that reproduces the folding probability best. As an input, a database of structures and their folding probabilities extracted uniformly with respect to p_{fold} is used. To determine the optimal weights for the different coordinates, a genetic neural network is employed. Applied to the conformational isomerization of the alanine dipeptide from $C_{7\text{eq}} \rightarrow \alpha_{\text{R}}$ in the presence of explicit water molecules, the procedure of Ma and Dinner identified a reasonable set of variables specifying the transition state [77].

Without relying on the folding probability, Best and Hummer [78] define a reaction coordinate by a weighted sum over the indicator functions of the contacts (value of the function is one, if the contact is formed and zero otherwise). The weights are iteratively refined such that the transition state structures are condensed into a single peak of the probability of being on a transition path. The transition paths and structures are determined by transition path sampling techniques or from equilibrium folding-unfolding trajectories. For a three-helix bundle test system studied by a Go-like Cα-model, the authors demonstrated that the identified transition state ensemble

matches the one obtained by traditional p_{fold} calculations. As pointed out by the authors, a careful definition of a contact is crucial to the successful application.

To construct a reaction coordinate for a particular transition between two free-energy basins, the string method of Ren et al. [79] and Maragliano et al. [80] can be used. The authors derive a differential equation for the minimum free-energy path (MFEP) based on a potential of mean force. Starting point for this derivation is the Langevin equation, i.e. Newton's equation of motion with a velocity proportional friction and a stochastic term mimicking the coupling to the temperature bath. Furthermore, the MFEP can be reparametrized analytically such that the MFEP-parameter is a reaction coordinate. To extend the reaction coordinate to the surrounding of the MFEP, a formula to calculate the region with equal value of the reaction coordinate is given. Applied to the isomerization of the dialanine peptide from the basin C_{7eq} to C_{7ax} the MFEP matches the predicted one. As the potential of mean force can be calculated also for a reduced set of coordinates, the MFEP and the reaction coordinate can be calculated for each set of coordinates. Thus, the influence of individual coordinates can be assessed. In the example of the dialanine peptide, the free-energy profile calculated from the $N - C_\alpha$ and $C_\alpha - C$ bond angles does not match the full-atomistic results considered here as the gold standard. If all backbone atoms are taken into account, the MFEP derived profile matches the reference.

Peters and Trout [81] extended the method of Best and Hummer [78] constructing a reaction coordinate from the weighted sum of the explicit values of the structural information and products thereof instead of indicator functions. A second remarkable difference is the usage of a statistical criterion (Bayesian Information Criterion) to determine the optimal number of parameters. The authors tested their procedure on synthetic potential energy landscapes and observed a good match of isoprobability lines and transition state region.

The forward flux sampling scheme (FFS) can be used to determine a reaction coordinate from a known order parameter [82]. The key idea behind FFS is to split the range of the order parameter into small intervals and to perform many short simulations connecting structures of both ends of a parameter interval. These data can be used to calculate a commitment probability p_{target} to the target basin by probabilistic means. The relation between structural properties (such as distances and angles) and the commitment probability is examined by identifying the optimal linear combination of structural properties by a least squares fit to p_{target}. Applying this method on a 64-mer model protein different nucleation scenarios were identified in agreement with experiments.

In essence, all reaction coordinates presented in this section rely on the calculation or approximation of the gold standard, i.e. the commitment/folding probability [73]. Numerous examples illustrate the success of this approach. But, if multiple disparate folding routes (e.g. for Beta3s) or intermediate states exist, the aforementioned descriptors may yield misleading results. The projection of the protein dynamics on networks allows the definition of reaction coordinates (cf. section 1.5.3 for details) that reliably maps the basins. This and other benefits from the projection on networks are discussed in the next section.

1.5 Analysis of *in silico* protein folding with networks

Networks allow the visualization of the interactions within high-dimensional data (for instance from sociology [83, 84], computer science [85], ecology [86] and the biochemistry of the metabolism [3, 87]) in low-dimensional graph representations. Apart from the ease of visualization, the application of networks provides immediate access to all results from graph theory, a mathematical discipline since decades. Results of graph theory relating building mechanisms and topological features can elucidate hidden structures in other fields. In *in silico* protein folding, the dynamical information of Monte Carlo or molecular dynamics (MD) simulations is mapped on networks [2, 88]. Applications such as the construction of reaction coordinates, the calculation of free-energy profiles and basins, and the determination of transition states are presented in the following sections.

1.5.1 Conformation space and equilibrium kinetic network

The conformation space network (CSN) is defined as follows: the nodes are coarse-grained structures and two vertices are linked when a transition between two structures occurs. A coarse-graining of the structures is necessary, because it is very unlikely to find the same structure twice in an MD simulation due to the limited sampling. Different coarse-graining procedures are discussed in the next section 1.5.2. If the simulation sampled the free-energy surface completely, the resulting CSN would fulfill the condition of detailed balance, i.e. the number of transitions $n_{A\to B}$ from nodes A to B are equal to the number of transitions in the opposite direction. The equilibrium kinetic network (EKN) is defined as the extrapolation of the CSN on perfect sam-

pling [89]. More precisely, the edges in the EKN are undirected and have a weight corresponding to the average number of transition in either direction.

1.5.2 Coarse-graining simulation

The root mean square deviation (RMSD) is the most popular metric to compare two structures. To calculate the RMSD, both structures are aligned to minimize the sum of the squared distances between the atom positions in both structures. The square root of the minimal sum is the RMSD value. Obviously, the C_α-RMSD (only C_α atoms are considered) does not discriminate between two structures with identical C_α-positions, but reoriented side chains (rotations of side chains may be the crucial difference between active and inactive conformation). Therefore, the choice of the subset of atoms used for comparison is important and depends on the application.

To identify clusters of similar structures in a long MD trajectory, the leader algorithm is employed (by this approach expensive all-to-all evaluations can be omitted). The first cluster has the first snapshots as a representative. Subsequently, for each snapshot at time t a similar structure is sought for (similarity is defined by the metric between snapshot and representative structure and the cutoff on it). If no matching representative is found, a new cluster with this snapshot as representative is formed. To maximize the extraction of local kinetic information, the comparison starts with the last snapshot before t and moves backwards in simulation time. The application of the leader algorithm is not restricted to RMSD, but also metrics on dihedral angles can be used [90].

A more coarse-grained approach to cluster structures relies on secondary structure detection. A standardized procedure [91] assigns secondary structure elements to individual residues according to backbone dihedral angles and hydrogen bonding patterns. Since the secondary structure elements are defined by few backbone contacts, a large variety of different structures with distinct side chain orientations are assigned to the same secondary structure.

1.5.3 Reaction coordinates from networks

Krivov and Karplus [89] demonstrated that the relative partition function Z_A/Z is a reaction coordinate. To calculate Z_A/Z for a given conformation, the smallest set A of nodes with weight Z_A (number of times a node in A is visited) in the EKN (cf. section 1.5.1) is determined such that the number of transitions crossing the boundary of A (Z_{AB} in Fig. 1.4) is minimal and both the initial and the given structure are in nodes of A. The value of Z_A/Z is then calculated as the fraction of the weight of A and the total number of

snapshots Z. By minimizing the number of A-boundary crossings, the free-energy barrier between the nodes in A and the remainder is maximized [89]. This definition of a barrier is independent of any other reaction coordinate and therefore preserves the barrier height under change of the projecting reaction coordinate. Furthermore, the free-energy basin of the initial structure is reliably represented, because for each value of Z_A/Z the highest barrier is taken into account. An accurate calculation of Z_A/Z is computationally very expensive. Therefore, two procedures have been suggested for approximation - *pfold* [89] and *mfpt* [99]. To this end, the "folding probability" *pfold* and the "mean first passage time" *mfpt* are calculated analytically for a random walker on the EKN. The different partitions of the network are given by different cutoffs on the *pfold* or *mfpt* values, respectively (Fig. 1.4). Notably, as *pfold* and *mfpt* both approximate the reaction coordinate Z_A/Z, they are both reaction coordinates as well. Examples for the successful application of these techniques are given in the next section.

1.5.4 Identifying states and free-energy basins of proteins

In the following section, four approaches to identify free-energy states and basins are discussed. Common to all procedures is that they rely on network technology.

Kinetic grouping analysis (KGA)

Conformations in the same free-energy basin interconvert more rapidly than structures from different free-energy basins [92]. Muff and Caflisch exploited this observation to identify free-energy basins of Beta3s, a 20-residue peptide with a native three-strand antiparallel β-sheet conformation [93]. To account for limited sampling, the structures recorded in an MD trajectory are coarse-grained. After coarse-graining, the clusters are merged if the probability p_{comm} of a transition between contained structures within a given commitment time τ_{comm} exceeds 0.5 (Fig. 1.2). Each set of pairwise merged clusters is considered to be one free-energy basin. The KGA procedure successfully elucidated the minor changes in the free-energy surface of Beta3s upon mutation of tryptophan at position 10 to valine [93]. Furthermore, KGA identified for a small helical peptide distinct folding routes and on-pathway traps that provided a mechanistic explanation for the stretched-exponential folding kinetics observed experimentally [43, 94]. Crucial to a successful application of the KGA procedure is that the interconversion time for structures within the same basin is significantly faster than the inter-basin transition. This

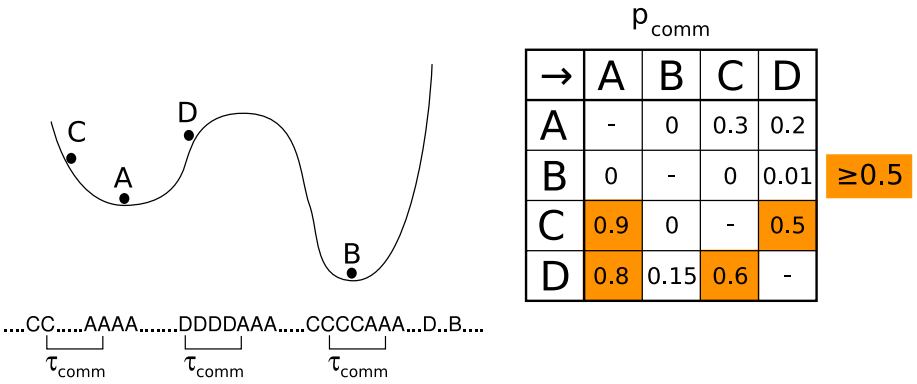

Figure 1.2: Illustration of KGA procedure (cf. section 1.5.4). Nodes A,C and D interconvert rapidly, because they belong to the same basin. In consequence, the probabilities p_{comm} of C and D to commit to the bottom node A within τ_{comm} are greater than or equal to 0.5. Therefore, these nodes are grouped by KGA. The node B is isolated, since a fast relaxation between B and the other nodes is prevented by the high barrier in between.

assumption is fulfilled well for mainly enthalpically stabilized basin, but not for very entropic regions. A direct extraction of barrier heights and transition rates is not possible. An advantage of the KGA procedure is that the determination of all free-energy basins requires only one calculation of the commitment probabilities.

Potential Energy Disconnectivity Graph (PEDG) and Transition Disconnectivity Graph (TRDG)

Potential energy (PEDG) [95] and Free-energy/Transition Disconnectivity Graphs (TRDG) [96] provide a two-dimensional representation of the potential energy and the free-energy surface, respectively. In both representations, the local minima of potential or free energy are displayed as leaves of a tree at a height reflecting their potential energy or free energy (Fig. 1.3). The branching points connecting two minima are placed such that their height corresponds to the height of the lowest barrier in between. Multiple schemes identify the local minima and saddle points of the potential energy landscape in the calculation of the PEDG [95]. These approaches are not discussed in detail here, because the study of the potential energy landscape provides only a limited picture of the processes involved in protein folding [19–22]. Therefore, Krivov and Karplus [92,96] modified the PEDG method [95] to analyze the free-energy landscape. The free-energy difference between two nodes is calculated on the EKN (cf. section 1.5.1) and given by the maximal number of transitions between them. According to the Ford-Flukerson [97] theorem this "maximal flow" is equivalent to the value of the mincut between the two

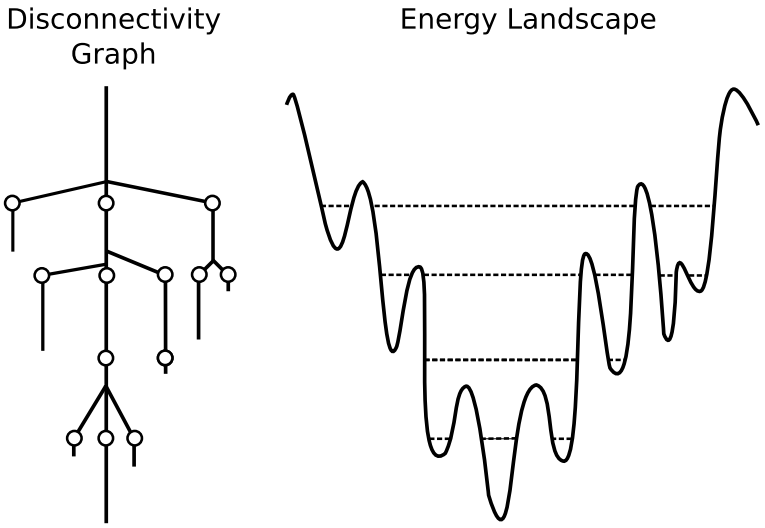

Figure 1.3: Example for a Disconnectivity Graph. Shown is the disconnectivity graph (left) of an artificial potential (right). The construction is described in section 1.5.4 or [96].

nodes. The mincut of two nodes in a network is the smallest sum of the capacities of edges that split the network into two parts upon removal with one node in each part. With the help of the Gomory-Hu theorem [98] all pairwise free-energy differences can be derived from the free-energy differences of N specific pairs. The rest of the free-energy disconnectivity graph calculation is identical to the TDRG calculation. Applied to the β-hairpin in protein G the TRDG displays four clearly pronounced free-energy minima indicating that the denatured state of the β-hairpin contains multiple basins [92].

By cutting the resulting trees at a height c (recall that the height in the tree corresponds to the potential or free-energy of the corresponding barrier or minimum), multiple subtrees are formed. They gather minima separated by barriers smaller than c. Varying the cutting height c allows to study the potential energy or free-energy surface at different levels of "resolution". In this respect, the cutting height of the TRDG approach has a similar function like the commitment time τ_{comm} for the kinetic grouping approach discussed earlier in the section.

Both the kinetic grouping approach (KGA) and the disconnectivity graphs merge structures not according to structural similarities, but according to their proximity in equilibrium dynamics. The disconnectivity graphs allow a direct extraction of barrier heights and therefore of interconversion rates. In contrast to KGA, an intrinsic determination of the optimum cutting height is not straight forward.

Cut-based free-energy profile

Cut-based free-energy profiles (cFEP) are one-dimensional projections of the free-energy surface explored from a given target structure. Remarkably, cFEPs preserve the barrier heights upon projection [89]. Furthermore, the free-energy basin containing the target node is reliably represented by all nodes in the cFEP to the left of the first barrier (cf. section 1.5.3). For each basin to detect, an additional cFEP has to be calculated, because the intervals of the progress variable of different free-energy basins might overlap on the right of the first cFEP-barrier. To calculate a cFEP from a given target node A, the EKN is splitted for each value of Z_A/Z into two parts $\mathcal{A}$ and $\mathcal{B}$ such that the number of transitions Z_{AB} across the boundary of $\mathcal{A}$ is minimized, the target node is in $\mathcal{A}$ and the number of times nodes in $\mathcal{A}$ are visited is equal to Z_A (Z is the number of snapshots). Henceforth, the point $(Z_A/Z, -k_BT\log(Z_{AB}))$ is added to the profile. This procedure is computationally very expensive. Therefore, two approximations have been suggested: *pfold* [89] and *mfpt* [99] (cf. section 1.5.3). The desired subsets (weight Z_A with minimal boundary crossings Z_{AB}) are approximated well by the sets of all snapshots with an alternative progress variable below a given threshold (Fig. 1.4). For the β-hairpin of protein G and analytical test potentials, the folding probability *pfold* (on the EKN) showed a good agreement between approximated and correct cFEP. The *mfpt*-based approach was employed to examine the free-energy landscape of Beta3s (three antiparallel β-strands peptide with 20 residues) [99]. The comparison of the three procedures (exact cFEP, *pfold*-cFEP and *mfpt*-CFEP) illustrated the equivalence of the three approaches, because the native states identified by the three procedures match to more than 99%.

To identify free-energy basins with cFEPs, the most visited node is chosen as first target for a cFEP. All snapshots to the left of the first peak of the cFEP belong to the same free-energy basin as the target node (cf. section 1.5.3). For the identification of a second basin, the most visited node outside the first basin is chosen and a second cFEP with this new target node is calculated. The second basin is formed by all nodes left to the first barrier of the second cFEP. This procedure is repeated until the cFEP displays no clear barrier any more. Alternatively, the different target nodes can be selected by choosing nodes in the various minima behind the first barrier in the initial cFEP [99]. This procedure has been applied to the identification of the free-energy states of Beta3s [99] (Fig. 1.5). Remarkably, the free-energy states identified by the KGA procedure [93] and the cFEP approach match to more than 96%. Furthermore, the cFEP approach was able to resolve subbasins in the free-energy basins identified by KGA. In comparison to KGA, no external

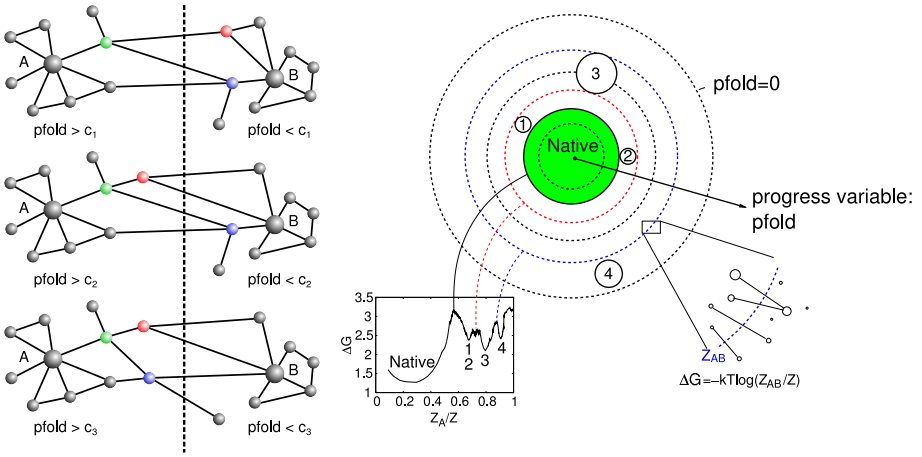

Figure 1.4: *pfold* Approximation to Cut-based Free-energy Profile (*pfold*-cFEP). The left hand side illustrates the role of *pfold* in the construction of the different sets A used in the reaction coordinate Z_A/Z. The plot on the right hand side exemplifies the overlap of non-native basins in the cFEP on the right of the first barrier. Details are given in section 1.5.4 or [89].

parameter like the commitment time is needed for the cFEP approach.

Network clusterization

Structures within the same free-energy basin interchange rapidly in the equilibrium dynamics (fundamental idea of KGA, earlier in this section). Thus, the nodes in the conformation space network (CSN) representing structures in the same free-energy basin should be substantially more connected among each other than to the other nodes. A vast collection of clustering algorithms (reviewed in more detail in section 1.7) can be used to detect such communities. In [90], Gfeller et al. compared the free-energy basins of the di-alanine peptide identified by three clustering algorithms on the CSN, where the structures are coarse-grained according to the dihedral angles. Tested are the MCL algorithm (Random walker with aging) [100], the Potts model clustering [101] and the greedy algorithm for modularity optimization [102]. Both the greedy and the Potts model algorithm merge multiple free-energy basins into one. This tendency to form too large communities was reported also for other applications [103, 104]. A good agreement with the reference basins is found for the MCL algorithm. The aging parameter in the MCL algorithm is able to tune the resolution in the detection of the free-energy basins. Noteworthily, existing community detection algorithms have been devised and better algorithms have been published. Nevertheless, it might be interesting to check the fundamental hypothesis (nodes in the same free-

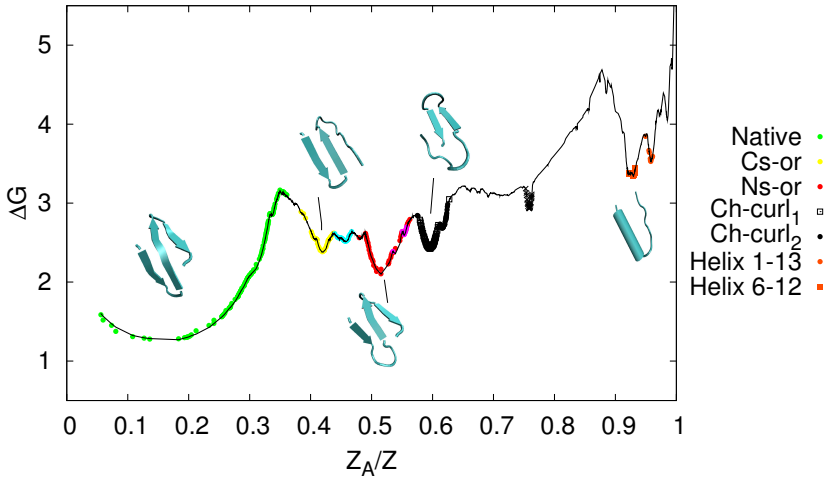

Figure 1.5: cFEP of Beta3s. The color of the symbols representing nodes with more than 100 snapshots indicates the corresponding basin affiliation. This figure is adopted from [99].

energy basin form a community) before testing other community detection schemes.

1.5.5 Identifying transition states

The folding behavior of a protein is determined not only by the free-energy minima, but also by the form of the transition state. A naive picture of the transition state is the ensemble of structures on the top of the barrier, separating one free-energy minimum from the others. More rigorously, the transition state is defined as the ensemble of structures with a folding probability $p_{\text{fold}} = 0.5$ [4, 73].

Without the application of network technology, the transition state structures can be identified using the reaction coordinates discussed in section 1.4. On the equilibrium kinetic network (EKN), a folding probability *pfold* can be defined as well (cf. section 1.5.4 or the original publication [89]). It approximates well the progress variable Z_A/Z [89] used to parametrize the cut-based free energy profile (cFEP). Therefore, the nodes close to the top of the first barrier in a cFEP should correspond to the transition state structures. Muff and Caflisch examined the transition state for Beta3s [105]. Structures are extracted uniformly distributed with respect to the reaction coordinate Z_A/Z of the cFEP calculated from the native structure. A complete p_{fold} calculation [73] is performed for each extracted structure by starting many simulations with different initial velocities and calculating the probability of

a trajectory to fold before unfolding. A comparison of the p_{fold} values with the location on the cFEP reveals that the structures with $p_{\text{fold}} = 0.5$ are located on the top of the barrier. In vicinity of the first barrier, a sudden transition of the average p_{fold} values from $p_{\text{fold}} \approx 1$ on the left of the barrier to $p_{\text{fold}} \approx 0$ on the right of the barrier can be observed. This consistency fortifies the faithful representation of the free-energy basin around the target node.

1.6 Förster Resonance Energy Transfer - a tool to study protein folding in vitro

Förster Resonance Energy transfer (FRET) is a promising technique to monitor an intermolecular distance of a single molecule with high time resolution [106, 107]. The attachment position of the donor and acceptor chromophore determines the intramolecular distance monitored. In an ideal experiment, the donor chromophore is excited by a laser (Fig. 1.6). This excited donor state can either decay to the ground state by the emission of a photon or transfer non-radiatively the excitation energy to the acceptor chromophore (the probability of this transfer depends on the distance between the two chromophores). For a suitably chosen pair of donor and acceptor chromophores, the wavelength of the emitted photon indicates whether a transfer has happened or not. Thus, the color sequence and the arrival times of the detected photons contain information about the instantaneous distance between the two chromophores. In fundamental contrast to ensemble techniques, FRET experiments monitor a single molecule and thus do not determine ensemble averaged quantities. For instance, distinct folding pathways can be observed for individual molecules. The time resolution of a FRET experiment is limited by the lifetime of the excited chromophore state (under the assumption of a fast transfer process compared to the lifetimes). For currently used chromophore systems, the life times are on the order of several nanoseconds. Other effects such as non-radiative transitions, low detection efficiency or dead times of detectors can lower the time resolution. Nevertheless, recent correlation experiments on DNA molecules reported a nanosecond time resolution [108]. As a further advantage, FRET experiments can be performed at very low protein concentrations. Therefore, spurious effects from protein aggregation can be significantly lowered.

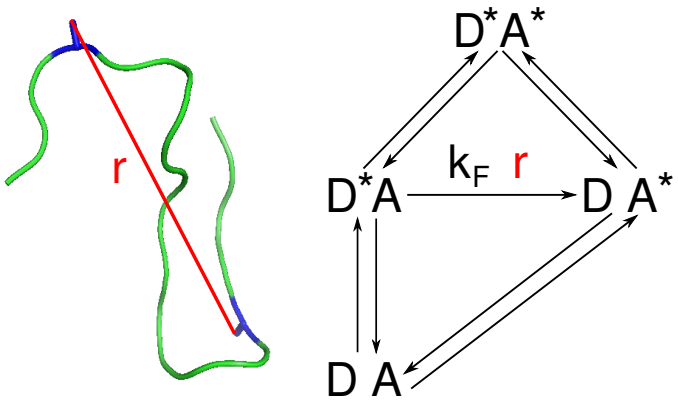

Figure 1.6: Fundamental Idea of FRET experiment. The diagram on the right displays the different states of donor (D) and acceptor (A) chromophore. The transfer of excitation from the donor to the acceptor chromophore is non-radiative and its rate $k_F(r)$ is distance dependent (cf. section 1.6). On the left hand side, the distance between the C_β-atoms of residues 4 and 16 in Beta3s is illustrated.

1.6.1 Bridging Experiment and Theory

In theoretical models such as molecular dynamics simulations, a time series for a single intramolecular distance can easily be extracted. In a FRET experiment, the measured photons originate from a stochastic process where the monitored distance influences the rate of the excitation transfer between donor and acceptor chromophore only. One strategy to bridge theory and experiments is to infer a distance trajectory from the photon trajectory. Schröder and Grubmüller [109] devised a maximum likelihood approach to calculate the probability distribution for the distance at an arbitrary time point in the measurement. This procedure was tested on a 10 ms burst with 230 photons recorded in a FRET experiment of syntaxin-1a. In a self-consistency test (comparing the distance trajectory inferred from a simulated and from the measured photon trajectory), the distance trajectories matched in the one-sigma interval. Watkins and Yang [110] suggest a procedure to determine the change points of the emission frequency in the limiting case of few photons detected between two change points. The key mathematical tool is a generalized likelihood ratio test. On synthetic data, Watkins and Yang can detect the correct change points with more than 90% certainty for only 20 photons emitted between two changes and the intensity of the two populations differing by a factor of five. However, both procedures infer a distance time series from the experimentally observed data.

Gopich and Szabo developed a new formalism [111–114] to answer the inverted question: Given an effective potential for the dynamics of the chro-

mophore distance, what would the corresponding photon trajectory look like? This question is addressed by analytically solving the underlying rate and diffusion equations. Furthermore, the authors reinterpret the solutions in terms of a field theory. By this analogy, the photon detection processes can be expressed as a perturbation to the unmonitored system. The field theory formalism can be utilized to find simple expressions for statistical moments of the interphoton time and number of photons distributions.

1.6.2 What can be learned from a single distance?

Compared to a protein's hundreds of degrees of freedom, a single distance appears little information. As not all degrees of freedom are crucial to monitor the folding process, a single distance time series might contain sufficient information to characterize free-energy basins. Baba et al. [115] tested their suggested simple procedure to extract free-energy basins of the BLN model, a simplified protein model with 46 "residues" built from three types of building blocks ("hydrophilic", "hydrophobic" and "neutral") natively folded in a β-barrel structure with four strands. The key idea of the procedure by Baba et al. is that the distance distribution in a short trajectory piece is characteristic for the free-energy basin. The different free-energy basins are determined by successively excluding those trajectory parts that look least like the distance distribution of the non-excluded snapshots. This method identifies four basins for the BLN-model that match the results of the TRDG method (discussed in section 1.5.4).

Li et al. [116] infer a state space network (SSN) from one-dimensional signal data. A state is defined by the set of time points with identical sequence of coarse-grained states for a short period around the time point considered. Wavelet technology with Haar's basis is used to coarse-grain the single distance time series and allows to tune the time resolution by the level of approximation. Applied on electron transfer experiments of the Fre/FAD complex, the identified SSN predicts an autocorrelation function of the fluorescence lifetime fluctuation that reproduces the observed dynamics on different time scales correctly.

The idea of Baba et al. that the short-time distribution of an observable is characteristic of a free-energy basin is enhanced in chapter 2. The FESST procedure presented there constructs a network of transitions akin to the equilibrium kinetic network (cf. section 1.5.1) for a molecular dynamics simulation comparing the signal distribution in short-time windows. One crucial point is that FESST maximizes the information of local kinetics extracted by merging the windows closest in time and fulfilling the similarity criterion. Due to the similarity of the network of transitions and the equilibrium ki-

netic network (EKN), the free-energy states are determined as from the EKN by the *pfold*-cFEP procedure (details in section 1.5.4). As application the distance trajectory between two side chain C_β-atoms of Beta3s, a 20-residue peptide with a native three β-sheet topology, is examined. FESST determines the native basin to 96% accuracy covering 95% of all native snapshots. As gold standard, the classification based on full-atomistic simulations in [99] is used. The folding-unfolding barrier is determined correctly (difference of only two percent). Furthermore, three non-native basins can be detected with high reliability. FESST can analyze as well other scalar signals. As an illustration, FESST is applied on the time series of FRET efficiencies derived from simulated photon trajectories. Even for low emission rates, the native state can be identified.

1.7 Further applications of networks

In the precedent sections, some applications of networks in the analysis of the protein/peptide dynamics are reviewed. Further application fields for networks in biochemistry are: study of metabolic pathways, protein-protein interaction data, gene expression networks. In general, the intrinsic structure of these large-scale data sets is unknown. Often, an organization into different functional modules, whose elements are tightly connected among each other, but loosely to elements of other modules, is assumed. To identify these clusters, a large number of different approaches has been developed in the last decade.

1.7.1 Community detection by generic procedures

The most elementary procedure for community identification is hierarchical clustering [117]. In each algorithm step, the pair of data points that is most similar in terms of the considered metric is merged. A classical application is microarray data, in which the activity profiles of different organisms have to be compared. This clusterization technique considers each expression profile as a node in a network with the similarity of two profiles as the capacity of the connecting edge. Hierarchical clustering is an agglomerative algorithm forming new communities in each step by merging two old ones. An example for a divisive algorithm is the betweenness clustering algorithm introduced by Girvan and Newman [118]. For each edge, a property called betweenness is calculated. It measures the number of shortest paths between any two nodes of the network containing the considered edge. In each algorithm step, the edge with the highest betweenness is removed. The rationale be-

hind this procedure is that inter-community edges in a network with clear community structure (many tightly connected subgraphs linked with few inter-community edges) are part of most shortest paths connecting nodes of different modules. Therefore, these inter-community edges will have a high betweenness and are removed early in the algorithm course. As a biological application, the betweenness clustering algorithm is applied to the food web of maritime organisms living in the Chesapeake Bay [119]. Apart from few unassigned species the clusterization algorithm splits the network into two groups that correspond well to the classification into benthic (living predominantly close to the bottom of the bay) and pelagic (living close to the surface) species. A major draw-back of both approaches is that the communities are defined according to the cutting height of a tree.

1.7.2 Optimization of a cost function to detect inherent structures

To relief this ambiguity of the cutting height, a cost function is introduced to measure the quality of the partition. The best clusterization is then the partition maximizing the assessment function. The quality function introduced first and broadly used is the modularity [120]. This cost function compares the fraction of intra-community edges with its expectation value for a random network with the same degree distribution. Although more objective in definition, the communities can only be determined after the choice of an optimization strategy. A broadly used and fast strategy is the greedy algorithm [102]. At first, each node forms its own community. In each algorithm iteration, those two communities are merged that increase the modularity most upon amalgamation. In the huge network of books copurchased at the online store Amazon (each book is a node and two nodes are connected if these books were purchased together) containing 409 687 vertices and 2 464 630 edges, the greedy algorithm is able to identify communities of books with common topics. In other examples such as the conformation space networks (discussed in section 1.5.4), the greedy algorithm tends to form too large communities. For the conformation space network of the dialanine peptide, several free-energy basins are merged. A more generic procedure called simulated annealing is used by Guimerà and Amaral to determine functional modules of metabolic networks [3]. The authors conclude that metabolites can adopt only seven distinct roles in a metabolic network. Remarkably, metabolites that participate in few reactions, but connect different modules are more conserved among the twelve species studied than hubs predominately linking nodes of the same module. In each iteration of

simulated annealing, a small part of the network is changed according to one prescription in a predefined set called move set. The change is accepted according to a Metropolis criterion (cf. section 1.3). Depending on the choice of the move set, the efficiency, accuracy and speed can be changed significantly. In most cases, one is altered on the expenses of the others.

A recent approach combining high accuracy and low running time is the multistep-greedy-vertex-mover algorithm [103, 104]. In a first step, a modification of the greedy algorithm is applied. The motivation to devise the existing greedy algorithm was to remedy the problem of excessive aggregation without sacrificing the low running time. To achieve this goal, multiple pairs are merged in each algorithm step. The optimal number of pairs to merge concurrently has to be fine-tuned by trial and error. But, there is strong evidence that the optimal number is in close vicinity of multiples of $0.25\sqrt{L}$ with L the total edge weight [104]. The results of the multistep greedy algorithm are afterwards refined by the "vertex mover" algorithm. The latter tests for each vertex (in the order of increasing number of connections) whether the reassignment of the considered node to a neighboring community yields an improvement of the cost function. This algorithm combination outperforms commonly used modularity optimization techniques on six out of seven benchmark networks. Furthermore, a single algorithm run has the lowest running time expectation among all other published procedures and similar running times compared to the greedy algorithm [121].

Modularity is not the only cost function to assess the quality of a network partition. Specialized on the detection of small communities is the localized modularity [122]. In comparison to modularity with a global view, localized modularity assesses the quality of the splitting only locally by considering for each community only the adjacent clusters. A different balance between the importance of intra-community links and the pressure for reduction of inter-community links can be reached by employing Potts models [101]. Approaches based on modularity and the Potts model are endowed with a resolution limit [123, 124], i.e. communities below a certain size are not identified individually but merged with one of the surrounding communities. To detect the individual modules, each community has to be considered as an autonomous network and has to be clustered individually.

Bibliography

[1] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5:101–113, 2004.

[2] F. Rao and A. Caflisch. The protein folding network. *J. Mol. Biol.*, 342:299–306, 2004.

[3] R. Guimerà and L. A. N.Amaral. Functional cartography of complex metabolic networks. *Nature (London)*, 433:895–900, 2005.

[4] A. Caflisch. Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.*, 16:71–78, 2006.

[5] C. Levinthal. Are there pathways for protein folding. *J. Chim. Phys.*, 65:44, 1968.

[6] R. Zwanzig, A. Szabo, and B. Bagchi. Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.*, 89(1):20–22, 1992.

[7] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. U.S.A.*, 89:8721–8725, 1992.

[8] A. Sali, E. Shakhnovich, and M. Karplus. How does a protein fold? *Nature (London)*, 369:248–251, 1994.

[9] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai. Navigating the folding routes. *Science*, 267:1619–1620, 1995.

[10] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 21:167–195, 1995.

[11] P. G. Wolynes. Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci. U.S.A.*, 93:14249–14255, 1996.

[12] M. Karplus. The Levinthal paradox: yesterday and today. *Fold. Des.*, 2:S69–S75, 1997.

[13] K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels. *Nat. Struct. Biol.*, 4:10–19, 1997.

[14] S. Arrhenius. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Z. Phys. Chem.*, 4:226–248, 1889.

[15] H. Eyring. The Activated Complex in Chemical Reactions. *J. Chem. Phys.*, 3:107, 1935.

[16] H. A. Kramers. Brownian Motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284, 1940.

[17] P. Faccioli, M. Sega, F. Pederiva, and H. Orland. Dominant pathways in protein folding. *Phys. Rev. Lett.*, 97:108101, 2006.

[18] A. van der Vaart and M. Karplus. Minimum free energy pathways and free energy profiles for conformational transitions based on atomistic molecular dynamics simulations. *J. Chem. Phys.*, 126:164106, 2007.

[19] A. R. Dinner, A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.*, 25:331–339, 2000.

[20] L. Mirny and E. Shakhnovich. Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.*, 30:361–396, 2001.

[21] V. Daggett and A. Fersht. The present view of the mechanism of protein folding. *Nat. Rev. Mol. Cell Biol.*, 4:497–502, 2003.

[22] P. G. Wolynes. Energy landscapes and solved protein-folding problems. *Philos. Transact. R. Soc. A*, 363:453–64; discussion 464–7, 2005.

[23] R. Bonneau, C. E. M. Strauss, C. A. Rohl, D. Chivian, P. Bradley, L. Malmström, T. Robertson, and D. Baker. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.*, 322:65–78, 2002.

[24] R. Bonneau, I. Ruczinski, J. Tsai, and D. Baker. Contact order and ab initio protein structure prediction. *Protein Sci.*, 11:1937–1944, 2002.

[25] P. Bradley, D. Chivian, J. Meiler, K. M. S. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. E. M. Strauss, and D. Baker. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, 53:457–468, 2003.

[26] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[27] N. Metropolis and S. Ulam. The Monte Carlo method. *J. Am. Stat. Assoc.*, 44:335–341, 1949.

[28] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equation of State Calculation by Fast Computing Machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[29] Z. Li and H. A. Scheraga. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.*, 84:6611–6615, 1987.

[30] V. Brodski, R. Peschar, and H. Schenk. A Monte Carlo approach to crystal structure determination from powder diffraction data. *J. Appl. Cryst.*, 36:239–243, 2003.

[31] S. Banu Ozkan and H. Meirovitch. Conformational search of peptides and proteins: Monte Carlo minimization with an adaptive bias method applied to the heptapeptide deltorphin. *J. Comp. Chem.*, 25:565–572, 2004.

[32] I. Newton. Philosophiae naturalis principia mathematica. page 530, 1726.

[33] B. R. Brooks, R. E. Bruccoler, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1982.

[34] W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[35] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.*, 118:11225–11236, 1996.

[36] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[37] J.E. Lennard-Jones. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proc. R. Soc.*, 106:463–477, 1924.

[38] W. Pauli. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Z. Phys.*, 33:765, 1925.

[39] E. Paci and M. Karplus. Unfolding proteins by external forces and temperature: the importance of topology and energetics. *Proc. Natl. Acad. Sci. U.S.A.*, 97:6521–6526, 2000.

[40] R. A. Böckmann and H. Grubmüller. Nanoseconds molecular dynamics simulation of primary mechanical energy transfer steps in F1-ATP synthase. *Nat. Struct. Biol.*, 9:198–202, 2002.

[41] C. D. Snow, L. Qiu, D. Du, F. Gai, S. J. Hagen, and V. S. Pande. Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.*, 101:4077–4082, 2004.

[42] D. A. C. Beck, G. W. N. White, and V. Daggett. Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *J. Struct. Biol.*, 157:514–523, 2007.

[43] J. A. Ihalainen, B. Paoli, S. Muff, E. H. G. Backus, J. Bredenbeck, G. A. Woolley, A. Caflisch, and P. Hamm. Alpha-Helix folding in the presence of structural constraints. *Proc. Natl. Acad. Sci. U.S.A.*, 105:9588–9593, 2008.

[44] W.E Reiher, III. *Theoretical Studies of Hydrogen Bonding*. PhD thesis, Department of Chemistry, Harvard University, Cambridge, MA, 1985.

[45] E. Neria, S. Fischer, and M. Karplus. Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, 105:1902–1921, 1996.

[46] A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[47] A. D. Mackerell Jr., M. Feig, and C. L. Brooks III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comp. Chem.*, 25:1400–1415, 2004.

[48] A. D. MacKerel Jr., N. Banavali, and N. Foloppe. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56:257–265, 2000.

[49] Y. Ueda, H. Taketomi, and N. Go. Studies of Protein Folding, Unfolding, and Fluctuations by Computer Simulation. II. A Three-Dimensional Lattice Model of Lysozyme. *Biopolymers*, 17:1531–1548, 1978.

[50] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[51] S. Nosé. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.*, 81:511–519, 1984.

[52] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[53] L. Boltzmann. Einige allgemeine Sätze über Wärmegleichgewicht. *Wiener Berichte*, 63:679–711, 1871.

[54] L. Boltzmann. Über die Natur der Gasmoleküle. *Wiener Berichte*, 74:553–560, 1876.

[55] D. Frenkel and B. Smit. *Understanding Molecular Simulation.* Academic Press; 2nd edition, 2001.

[56] V. A. Likié and F. G. Prendergast. Dynamics of internal water in fatty acid binding protein: Computer simulations and comparison with experiments. *Proteins*, 43:65–72, 2001.

[57] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[58] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins*, 35:133–152, 1999.

[59] P. Ferrara, J. Apostolakis, and A. Caflisch. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins*, 46:24–33, 2002.

[60] U. Haberthür and A. Caflisch. FACTS: Fast analytical continuum treatment of solvation. *J. Comp. Chem.*, 29:701–715, 2008.

[61] E. de Alba, J. Santoro, M. Rico, and M. A. Jimnez. De novo design of a monomeric three-stranded antiparallel beta-sheet. *Protein Sci.*, 8:854–865, 1999.

[62] G. Settanni, F. Rao, and A. Caflisch. Phi-value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl. Acad. Sci. U.S.A.*, 102:628–633, 2005.

[63] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.

[64] C. Dellago, P. G. Bolhuis, and P. L. Geissler. Transition Path Sampling. *Advances in Chemical Physics*, pages 1–78, 2003.

[65] P. G. Bolhuis. Transition-path sampling of beta-hairpin folding. *Proc. Natl. Acad. Sci. U.S.A.*, 100:12129–12134, 2003.

[66] H. Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev. E*, 52:2893–2906, 1995.

[67] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.*, 120:11919–11929, 2004.

[68] K. Hamacher and W. Wenzel. Scaling behavior of stochastic minimization algorithms in a perfect funnel landscape. *Phys. Rev. E*, 59:938–941, 1999.

[69] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[70] R. Zhou, B. J. Berne, and R. Germain. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U.S.A.*, 98:14931–14936, 2001.

[71] S. Muff and A. Caflisch. ETNA: Equilibrium Transitions Network and Arrhenius Equation for Extracting Folding Kinetics from REMD Simulations. *J. Phys. Chem. B*, 113:3218–3226, 2009.

[72] A. D. McNaught and A. Wilkinson. *IUPAC Compendium of Chemical Terminology.* Blackwell Science, 1997.

[73] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108:334–350, 1998.

[74] M. Karplus. Aspects of Protein Reaction Dynamics:Deviations from Simple Behavior. *J. Phys. Chem. B*, 104:11–27, 2000.

[75] S. S. Cho, Y. Levy, and P. G. Wolynes. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 103:586–591, 2006.

[76] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch. Estimation of protein folding probability from equilibrium simulations. *J. Chem. Phys.*, 122:184901, 2005.

[77] A. Ma and A. R. Dinner. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109:6769–6779, 2005.

[78] R. B. Best and G. Hummer. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. U.S.A.*, 102:6732–6737, 2005.

[79] W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *J. Chem. Phys.*, 123:134109, 2005.

[80] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.*, 125:24106, 2006.

[81] B. Peters and B. L. Trout. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.*, 125:054108, 2006.

[82] E. E. Borrero and F. A. Escobedo. Reaction coordinates and transition pathways of rare events via forward flux sampling. *J. Chem. Phys.*, 127:164101, 2007.

[83] W. W. Zachary. Information-Flow Model for Conflict and Fission in Small-Groups. *J. Anthropol. Res.*, 33:452 – 473, 1974.

[84] M. E. J. Newman. The structure of scientific collaboration. *Proc. Natl. Acad. Sci. U.S.A.*, 98 (2):404 – 409, 2001.

[85] Internet Network: undirected, unweighted network of the Internet at the Autonomous System level from data collected by the Oregon Route Views Project (http://www.routeviews.org/) in May 2001, where vertices represent Internet service providers and edges connections among them. The file reports the list of connected pairs of nodes.

[86] N. D. Martinez. Artifacts or Attributes? Effects of Resolution on the Little Rock Lake Food Web. *Ecological Monographs*, 61:367–392, 1991.

[87] H. Ma and A. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19:270–277, 2003.

[88] A. Scala, L. A. N. Amaral, and M. Barthelemy. Small-world networks and the conformation space of a short lattice polymer chain. *Europhysics Letters*, 55:594–600, 2001.

[89] S. V. Krivov and M. Karplus. One-dimensional free-energy profiles of complex systems: progress variables that preserve the barriers. *J. Phys. Chem. B*, 110:12689–12698, 2006.

[90] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. Complex network analysis of free-energy landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 104:1817–1822, 2007.

[91] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[92] S. V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.*, 101:14766–14770, 2004.

[93] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins*, 70:1185–1195, 2008.

[94] B. Paoli, M. Seeber, E. H. G. Backus, J. A. Ihalainen, P. Hamm, and A. Caflisch. Bulky Side Chains and Non-native Salt Bridges Slow down the Folding of a Cross-Linked Helical Peptide: A Combined Molecular Dynamics and Time-Resolved Infrared Spectroscopy Study. *J. Phys. Chem B*, 113:4435–4442, 2009.

[95] O. M. Becker and M. Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106:1495–1517, 1997.

[96] S. V. Krivov and M. Karplus. Free energy disconnectivity graphs: Application to peptide models. *J. Chem. Phys.*, 117:10894–10903, 2002.

[97] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Can. J. Math.*, 8:399, 1956.

[98] R. E. Gomory and T. C. Hu. Multi-terminal network flows. *Operations Research*, pages 551–570, 1961.

[99] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus. One-dimensional barrier-preserving free-energy projections of a beta-sheet miniprotein: new insights into the folding process. *J. Phys. Chem. B*, 112:8701–8714, 2008.

[100] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30:1575–1584, 2002.

[101] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.*, 93:218701, 2004.

[102] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, 2004.

[103] P. Schuetz and A. Caflisch. Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. *Phys. Rev. E*, 78:026112, 2008.

[104] P. Schuetz and A. Caflisch. Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Phys. Rev. E*, 77:046112, 2008.

[105] S. Muff and A. Caflisch. Identification of the protein folding transition state from molecular dynamics trajectories. *J. Chem. Phys.*, 130:125104, 2009.

[106] T. Förster. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der Physik*, 437(1-2):55–75, 1948.

[107] L. Stryer. Fluorescence Energy Transfer as a Spectroscopic Ruler. *Annual Review of Biochemistry*, 47(1):819–846, 1978.

[108] A. J. Berglund, A. C. Doherty, and H. Mabuchi. Photon statistics and dynamics of fluorescence resonance energy transfer. *Phys. Rev. Lett.*, 89:068101, 2002.

[109] G. F. Schröder and H. Grubmüller. Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer experiments. *J. Chem. Phys.*, 119:9920–9924, 2003.

[110] L. P. Watkins and H. Yang. Detection of intensity change points in time-resolved single-molecule measurements. *J. Phys. Chem. B*, 109:617–628, 2005.

[111] I. Gopich and A. Szabo. Fluorophore-quencher distance correlation functions from single-molecule photon arrival trajectories. *J. Phys. Chem. B*, 109:6845–6848, 2005.

[112] I. Gopich and A. Szabo. Theory of photon statistics in single-molecule Frster resonance energy transfer. *J. Chem. Phys.*, 122:14707, 2005.

[113] I. V. Gopich and A. Szabo. Single-molecule FRET with diffusion and conformational dynamics. *J. Phys. Chem. B*, 111:12925–12932, 2007.

[114] I. V. Gopich and A. Szabo. Photon counting histograms for diffusing fluorophores. *J. Phys. Chem. B*, 109:17683–17688, 2005.

[115] A. Baba and T. Komatsuzaki. Construction of effective free energy landscape from single-molecule time series. *Proc. Natl. Acad. Sci. U.S.A.*, 104:19297–19302, 2007.

[116] C. Li, H. Yang, and T. Komatsuzaki. Multiscale complex network of protein conformational fluctuations in single-molecule time series. *Proc. Natl. Acad. Sci. U.S.A.*, 105:536–541, 2008.

[117] J. H. Ward. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Ass.*, 58:236–244, 1963.

[118] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci U.S.A.*, 99:7821–7826, 2002.

[119] D. Baird and R. E. Ulanowicz. The Seasonal Dynamics of The Chesapeake Bay Ecosystem. *Ecological Monographs*, 59:329–364, 1989.

[120] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[121] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.*, 2005:P09008, 2005.

[122] S. Muff, F. Rao, and A. Caflisch. Local modularity measure for network clusterizations. *Phys. Rev. E*, 72:056107, 2005.

[123] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proc. Natl. Acad. Sci. U.S.A.*, 104:36–41, 2007.

[124] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész. Limited resolution in complex network community detection with Potts model approach. *Eur. Phys. J. B*, 56:41 – 45, 2007.

Chapter 2

Free energy surfaces from single-distance information

P. Schuetz, B. Schuler, and A. Caflisch

[submitted]

Free energy surfaces from single-distance information

Philipp Schuetz, Ben Schuler*, and Amedeo Caflisch*

Department of Biochemistry
University of Zürich, Winterthurerstrasse 190
CH-8057 Zürich, Switzerland
Phone: (+41 44) 635 55 21, FAX: (+41 44) 635 68 62
email: caflisch@bioc.uzh.ch, schuler@bioc.uzh.ch

*Corresponding authors

Abstract

We propose a method for determining basins and barriers of complex free energy surfaces (e.g., the protein folding landscape) from the time series of a single intramolecular distance. First, a network of transitions is constructed by clustering the points of the time series according to the short-time distribution of the signal. The transition network, which reflects the short-time kinetics, is then used for the iterative determination of individual basins by the minimum-cut-based free energy profile, a barrier-preserving one-dimensional projection of the free energy surface. The method is tested using the time series of a single C_β-C_β distance extracted from equilibrium molecular dynamics (MD) simulations of a structured peptide (20-residue three-stranded antiparallel β-sheet). Although the information of only one distance is employed to describe a system with 645 degrees of freedom, both the native state and the unfolding barrier of about 10 kJ/mol are determined with remarkable accuracy. Moreover, non-native conformers are identified by comparing long-time distributions of the same distance. To examine the applicability to single-molecule Förster resonance energy transfer (FRET) experiments, a time series of donor and acceptor photons is generated using the MD trajectory. The native state of the β-sheet peptide is determined accurately from the emulated FRET signal, which indicates that the method is useful for extracting free energy surfaces from single-molecule experiments.

Introduction

The thermodynamics and kinetics of a variety of complex systems, ranging from spin glasses to proteins, have been investigated by energy landscape theory in the 40 years since the publication of the seminal idea [1]. Peptides and proteins have a multidimensional and very complex potential energy surface with a large number of conformations of similar energy [2, 3]. Yet, fast folding is possible because of the natural selection of sequences that make the native (i.e., functional) structure a pronounced energy minimum [4]. Entropic contributions are relevant at physiological temperatures and therefore the *free* energy surface governs the thermodynamics and kinetics of polypeptide chains. In the past five years, new methods based on complex networks have been proposed to analyze free energy surfaces of folding [5, 6, 7, 8], which govern the process by which structured peptides or proteins assume their well-defined three-dimensional structure.

In view of the large number of microscopic folding pathways and the conformational heterogeneity in the denatured state, single molecule methods are a promising new approach to experimentally determine free energy surfaces [9]. One of the most versatile approaches, single molecule Förster resonance energy transfer (FRET), allows intramolecular distances and distance dynamics of individual protein molecules to be monitored [10, 11, 12, 13, 14, 15, 16]. Since distance distributions in different free energy states often overlap, the separability of the different basins is not straightforward. Baba and Komatsuzaki suggested an approach (termed BK procedure hereafter) to extract free energy basins from the time series of a single distance [17]. The BK procedure is able to resolve different basins even if the distance distributions overlap because the short-time behavior of the observable is considered. Applied to a simplified model of a protein with 46 beads of three types (hydrophobic, hydrophilic, and neutral), the authors identified four free energy basins, in good agreement with the free energy surface derived using the complete structural information of the reference simulation.

Here we present a procedure for the automatic determination of free energy

surfaces from single-molecule time series (FESST). First, an equilibrium transition network (ETN) is constructed by clustering individual time windows according to similarity in the short-time distribution of the signal, whose usage was inspired by the BK procedure [17]. The ETN is then used as the input for the minimum-cut-based free energy profile (cFEP) method, which is able to determine free energy basins and barrier heights [7]. The FESST parameters are optimized using an intrinsic cost function, the height of the unfolding barrier in the cFEP, which allows for complete automatization of the procedure.

The accuracy of FESST is assessed using molecular dynamics (MD) trajectories of the 20-residue peptide Beta3s [18, 19], whose sequence was designed to favor the three-stranded antiparallel β-sheet conformation, i.e., a double β-hairpin [20]. Beta3s has been shown to fold reversibly to the native structure determined by NMR [20] in MD simulations with the CHARMM polar hydrogen molecular mechanics potential energy function supplemented by a simple implicit solvent model [21]. In these simulations, Beta3s folds in about 0.1 μs and 8 μs at 330 K and 286 K, respectively [22]. Since multiple folding and unfolding events at the melting temperature of about 330 K can be simulated in less than a week (on a commodity processor), the free energy surface, as well as the folding pathways and mechanism of Beta3s have been investigated in detail [6, 18, 19, 23]. The complexity of the free energy surface of Beta3s [6] and its detailed characterization make it an ideal test system. FESST is able to extract the native basin of Beta3s and the height of the free energy barrier between folded and unfolded state correctly using only the time series of the distance between two C_β atoms. In addition, FESST identifies three subbasins in the denaturate state with high completeness and good accuracy. The application of FESST is not limited to distance information, but any type of one-dimensional signal can be used, if the signal has a characteristic range of values for each (sub)basin. Furthermore, FESST is assessed on the time series of FRET efficiencies from an emulated single-molecule experiment to analyze its applicability to experimental data. Evidence is provided that FESST identifies the native

basin with high accuracy and completeness even for a small number of detected photons.

Methodology

Free energy surface from single-molecule time series (FESST). FESST is a three-step procedure: construction of the ETN by clustering individual time windows using local kinetic information, identification of free energy basins by the cFEP approach, and removal of overlap from the non-native basins. The details of the three steps of FESST are presented in the next subsections while a schematic illustration is shown in Fig. 1.

Coarse-graining and equilibrium transition network (ETN). Each time window in the one-dimensional signal is assigned to a node of the ETN by the leader algorithm [24]. In the initialization step, the first time window is defined as the representative of the first node. At each successive time window t_n ($n > 1$), the short-time distribution of the single-molecule signal is compared with those of the previously visited representatives. To preserve the local kinetics (i.e., the actual dynamic evolution of the system), the comparison is carried out starting from the latest defined representative, i.e., by parsing the list of representatives in inverse chronological order. A new node is defined whenever the short-time distribution of t_n deviates by more than a given threshold from those of all the previously defined representatives. In this way, one obtains a time series of nodes and a corresponding sequence of transitions between nodes, which is used to construct the ETN (Fig. 1a,b).

Minimum-cut-based free energy profile (cFEP). Krivov and Karplus have exploited an analogy between the kinetics of a complex process and equilibrium flow through a network to develop the cFEP, a projection of the free energy surface that preserves the barriers [7] and can be used for extracting folding pathways and mechanisms from MD simulations [25]. The input for the cFEP calculation is the ETN, which is derived by the coarse-graining described above. For a node i in

the ETN the partition function is $Z_i = \sum_j c_{ij}$, where $c_{ij} = \frac{n_{ji}+n_{ij}}{2}$, and n_{ji} is the absolute number of transitions from node i to node j observed along the time series, so that c_{ij} are the entries of the symmetrized transition matrix that satisfies detailed balance. The transition probabilities can then be calculated as $p_{ij} = c_{ij}/\sum_k c_{kj}$. If the nodes of the ETN are partitioned into two sets $\mathcal{A}$ and $\mathcal{B}$, where set $\mathcal{A}$ contains the reference node A, then $Z_A = \sum_{i\in\mathcal{A}} Z_i$, $Z_B = \sum_{i\in\mathcal{B}} Z_i$, $Z_{AB} = \sum_{i\in\mathcal{A},j\in\mathcal{B}} c_{ij}$, and the free energy of the barrier between the two groups is $\Delta G = -kT\log(Z_{AB}/Z)$, where Z is the partition function of the full ETN. The progress coordinate then is the normalized partition function Z_A/Z of the reactant region containing the native node A, but other progress coordinates can be used, because the cFEP is invariant with respect to arbitrary transformations of the reaction coordinate [26]. The cFEP is calculated from the ETN in three steps: (1) The folding probability p_{fold} is calculated analytically for each node on the ETN by solving the system of transition rate equations [7, 25]. (2) Nodes are sorted by decreasing values of p_{fold}, and for each of these values the relative partition function Z_A and the cut Z_{AB} are calculated. (3) The individual points on the profile are evaluated as $[x = Z_A/Z$, $y = -kT\log(Z_{AB}/Z)]$ (Fig. 1c,d). The result is a one-dimensional profile that preserves the barrier heights between the free energy basins; given the barriers, the basins can be determined [7].

Iterative determination of free energy basins. The most populated node is used to isolate the first basin by the cFEP approach. In this way, the native basin is usually isolated first, so that the corresponding cFEP shows the unfolding barrier (Fig. 1c). For the remaining basins, the procedure is the same, except that the most populated unassigned node is used as a reference (Fig. 1d). All nodes to the left of the cut at the first barrier make up the basin. Basins to the right of the first barrier are potentially overlapping, thus each basin requires a separate "exiting" profile [25]. Moreover, a FESST basin encompasses usually more than one of the actual free energy basins, because the short-time distribution of the single distance can be degenerate. To remove this overlap, the long-time distribution of the signal

in each time window assigned to the considered FESST basin is compared with the distribution of the signal in the entire basin identified previously (Fig. 1e). To exploit information complementary to the one used in the construction of the ETN, longer intervals of the time series around each time window are compared. Different subbasins are characterized by different ranges of the comparison metric (Fig. 1f), because two distinct free-energy basins differ in their similarity to a third one.

Application to atomistic MD simulations of β-sheet folding

MD simulations of Beta3s. A total simulation time of 20 μs at 330K was used for the FESST analysis. It has been shown previously that in MD at 330K Beta3s folds reversibly to the NMR conformation, irrespective of the starting structure; 23 of the 26 nuclear Overhauser effect constraints are satisfied [18, 19]. All MD runs and most of the analysis of the trajectories were carried out with CHARMM [27, 28]; the rest of the analysis was done with the program WORDOM [29]. The designed 20-residue peptide Beta3s [20] (Thr_1-Trp_2-Ile_3-Gln_4-Asn_5-Gly_6-Ser_7-Thr_8-Lys_9-Trp_{10}-Tyr_{11}-Gln_{12}-Asn_{13}-Gly_{14}-Ser_{15}-Thr_{16}-Lys_{17}-Ile_{18}-Tyr_{19}-Thr_{20}) was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field [27, 30] with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent [21]. More explicitly, the screening of the electrostatic interactions is approximated by the distance-dependent dielectric function $\epsilon(r) = 2r$, while the remaining solvation effects are approximated by replacement of the monopole moment of charged groups by strong dipole moments and a linear function of atomic SAS values. The latter requires only two surface-tension like parameters and takes into account both polar and apolar solvation effects by a negative (i.e., favorable) value of the surface-tension parameter for nitrogen and oxygen atoms, and a positive (unfavorable) value for carbon and sulfur atoms.

Intramolecular distance and metric used for coarse-graining. The time series of the C_β Gln_4 - C_β Thr_{16} distance is used in FESST, and the robustness of the results with respect to the choice of distance is discussed below. Two time windows $[t_1, t_1 + \tau]$ and $[t_2, t_2 + \tau]$ are grouped into the same node of the ETN if their distributions of the intramolecular distance (the short-term distribution) pass a Kolmogorov-Smirnov test [31], which checks whether two samples are picked from the same distribution. Each MD snapshot is used as a starting point of a time window, so that there are as many windows as coordinate frames along the MD trajectory. In other words, two successive windows are shifted by the MD saving interval of 20 ps. The time window τ is much shorter than the folding time, which is about 100 ns in MD simulations of Beta3s at 330 K [19]. The disparity of the two time windows is defined as the maximum difference of the cumulative distribution functions c_1, c_2 of the distance r (disparity $= \max_{r>0} |c_1(r) - c_2(r)|$). The test is passed if

$$\text{disparity} \leq \sqrt{\frac{2}{\mathrm{N}}} \cdot \zeta$$

with N the number of MD snapshots in each time window and ζ the acceptance cutoff that corresponds to a certain confidence level [31]. Note that the FESST results on Beta3s are robust with respect to the choice of N in the range $30 \leq \mathrm{N} \leq 250$ (i.e., 0.6 ns $\leq \tau \leq$ 5 ns) and ζ in the range $0.3 \leq \zeta \leq 1.5$ (Fig. S1). Values of $\tau = 2$ ns and $\zeta = 0.3$ are used in the following.

Native basin and unfolding barrier. The free energy basins of Beta3s have been determined previously by the cFEP procedure using information on all the 645 coordinates [25] (hereafter abbreviated as 645-coords cFEP). Since the full information of the peptide dynamics was taken into account, those free energy basins and barriers are used here as a reference for a critical evaluation of FESST and comparison with other approaches. The native basin of Beta3s is determined by FESST with remarkable accuracy (96% of the FESST native basin belongs to the native state as determined by the 645-coords cFEP) and completeness (95% of the native state of the 645-coords cFEP is captured by the FESST native basin).

Moreover, the FESST unfolding barrier has very similar height (10.7 kJ/mol) as the one obtained by the 645-coords cFEP (10.6 kJ/mol, Fig. 2a and Fig. S2). FESST performs much better than a null model (Fig. 2b), in which the native state is determined by a histogram-based free energy projection onto the single distance, because a single distance value is degenerate, i.e., conformations from different basins can have similar scalar values of the distance (Fig. 3a). Although both approaches make use of short-time distributions of the signal, FESST has two advantages with respect to the BK procedure [17]. First, FESST exploits the local kinetic information for the coarse-graining, while the BK procedure iteratively removes the time windows least similar to the distribution of the whole distance time series, thus ignoring the chronological order of the windows. Second, the optimal values of parameters required by FESST (size of the time window k and acceptance cutoff ζ used in the coarse-graining) can be determined automatically using the cFEP barrier height as cost function, because the most accurate determination of the native basin yields the highest barrier between it and the other basins [25]. Note that the parameter set yielding the highest barrier achieves the highest score (defined as the product of accuracy and completeness, Fig. 2c). Therefore, FESST yields a single data point in the accuracy vs. completeness plot (Fig. 2b), whereas the basins extracted by other procedures depend on the cutoffs chosen for their iterative refinement, so that it is not possible to automatically identify the optimal solution.

Robustness with respect to the choice of the intramolecular distance monitored. To investigate the influence of the choice of the residue pair monitored, each of the 154 C_β-C_β pairs was tested in FESST. Remarkably, for 32 of these pairs the native basin is identified with an accuracy greater than 80% and at the same time a completeness of more than 90% (Fig. 3b). Interestingly, the larger the separation along the sequence, the better the score. A notable exception is the 5-7 distance, which reflects the formation of the β-turn at the N-terminal hairpin. The distances yielding the best score are those between residues in β-strands 1

and 3 (top, left part of the matrix in Fig. 3b), which is likely to be a consequence of the β-sheet topology. Moreover, the C_β-C_β distances involving the N-terminal β-strand show a higher score than those involving the C-terminal β-strand, which is consistent with the higher structural stability of the C-terminal hairpin [18, 19]. In other words, the fully folded state can be better separated from non-native conformers (discussed below) by taking into account the N-terminal β-strand, because the C-terminal hairpin is folded correctly in the most populated non-native conformers.

Identification of non-native basins. The most populated node outside of the native basin is used as a reference to plot the cFEP profile for identifying the first non-native basin (termed $\tilde{B}_2$ in Fig. 1). Due to the degeneracy of the short-time distribution of the distance, multiple free energy basins may overlap on the cFEP. Such overlap can be removed by comparing the long-time distance distribution of each time window with the distance distribution in a previously identified basin. In practice, for each time window $[t_2, t_2 + T]$ in basin $\tilde{B}_2$, the distribution of the distance is compared with the histogram of the whole native basin. Time windows (length T $\approx$(10 to 20)$\cdot\tau$) larger than those used for the construction of the ETN are considered here for better statistics. The comparison consists of calculating the Kantorovich metric [32] between the two distributions (the area between the two cumulative histograms, Fig. 1e). Finally, each peak in the histogram of the Kantorovich values is assigned to a new subbasin (Fig. 1f). The window size T can be chosen by optimizing the separation of the different peaks in the Kantorovich histogram (Fig. S3).

With this procedure, the basin B_2 derived from $\tilde{B}_2$ corresponds to the 645-coords free energy basin Ch-curl$_1$ (curl-like conformation with folded C-terminal hairpin [25]) with 92% accuracy and 85% completeness. Further, the third FESST basin $\tilde{B}_3$ encompasses two free energy basins and can be split by comparing with the distance distribution in $\tilde{B}_2$. The free energy basin Ns-or$_1$ (N-terminal strand out of register and folded C-terminal hairpin [25]) can be extracted with 77% accuracy

and 68% completeness. The second subbasin detected in B_3 contains 56% of MD snapshots in Ch-curl$_2$ (curl-like conformation 2 [25]) covering 77% of these MD snapshots. These non-native conformers are stabilized mainly enthalpically [25]. Entropically stabilized conformations such as those in the "helical basin" and the "entropic region" [25] show a very broad distribution of distances (blue and gray curves in Fig. 3a). These broad distributions overlap strongly with those of other basins and therefore are distributed over multiple FESST basins. In other words, both $\tilde{B}_2$ and $\tilde{B}_3$ contain time windows of the entropic region that can be removed by the procedure illustrated in Fig. 1e,f. [1]

Application to an emulated FRET signal

To mimic FRET experiments, the sequence of states visited by a random walker in the Markov state model of the FRET process (inset of Fig. 4) is recorded (details in supporting information (SI)). In this model, the rate of energy transfer $k_F(r)$ between the two "virtual" chromophores depends on the inverse sixth power of the distance r between the C_β-atoms of Beta3s residues 4 and 16 as recorded along the MD trajectory. To improve the statistics at low emission rates, the photon time series is split into bins of 0.4 ns, i.e., 20 MD snapshots. FESST analyzes the time series of FRET efficiencies $E_{\mathrm{FRET}} = \frac{n_\mathrm{A}}{n_\mathrm{A}+n_\mathrm{D}}$ with n_A and n_D the number of acceptor and donor photons in the considered bin.

The effect of the number of photons per bin is studied by the variation of the excitation rate (Fig. 4). For comparability with experiments, we report the number of photons emitted during the folding time. The score of the native state detection increases with the average emission rate, and reaches a plateau of 85% accuracy and 88% completeness at an emission rate of about 5000 photons per folding time (Fig. 4), compared to 96% accuracy and 95% completeness obtained by applying FESST to the distance time series without binning. Importantly, the native state of Beta3s can be detected with 78% accuracy and 78% completeness from as few

[1] For the native basin, this step is not performed, because the overlap of the distance distributions is much smaller than for $\tilde{B}_2$ and $\tilde{B}_3$, and no basin for comparison is available.

as 1000 photons per folding time. This detection quality is obtained by comparing intervals of the time series of FRET efficiencies as long as 10 nanoseconds, which corresponds to about one tenth of the folding time of Beta3s in the MD simulation at the melting temperature. The detection quality depends only weakly on the size of each FRET bin (Fig. S4) and the length of the time series interval (Fig. S5).

Resolution limits in the analysis of FRET experiments. It is useful to investigate the resolution limit of FESST using a simple model (Fig. S11). The time evolution of the monitored signal is given by Langevin dynamics of a particle in a one-dimensional potential. To model a two-state system, the potential is switched with a constant rate between two harmonic wells (Fig. S11 a,b,c). The sequence of emitted photons is determined by a Gillespie-type simulation [33]. The time series of FRET efficiencies is derived from the binned photon sequence and analyzed by FESST as described for Beta3s. Remarkably, the accuracy of FESST is always higher than the null model even for very small separations of the minima (Fig. S11 d,f,g). FESST can discriminate the minima based on the curvature alone, albeit with a relatively large number of detected photons required (cf. Fig. S11 and Fig. S12). Although the simple model does not reflect the complexity of a multidimensional system, the present results indicate that the reliable operation of FESST requires the detection of 100 to 1000 photons while the system stays in one free-energy state. In real single-molecule FRET experiments, 100 to 1000 photons can be detected in about one to ten milliseconds. Accordingly, we expect FESST to be a suitable approach for determining the properties of free energy surfaces of molecules that exhibit dynamics in this range or slower, thus covering a large part of the biologically important time scales [34, 35, 36, 37].

Concluding discussion

FESST (free energy surface from single-molecule time series) is a method for determining free energy basins and barriers from the time evolution of a scalar observable. The accuracy and range of possible applications of FESST have been

investigated using scalar time series derived from atomistic MD simulations of the reversible folding of a structured peptide. First, FESST was applied to the time series of a single interresidue distance of Beta3s, a 20-residue peptide with native three-stranded β-sheet topology. The native state of Beta3s, three subbasins in the denaturate state, and the free energy barrier for unfolding can be determined with high accuracy. Importantly, FESST is robust under the choice of the residue pair. In fact, each of about 20% of the 154 pairs of C_β-C_β distances can be used in FESST for determining the native state of Beta3s, and in particular distances between residues in β-strands 1 and 3 are optimal. Furthermore, the basin assignment by FESST is robust under change of the parameters used for coarse-graining, which can be determined self-consistently.

In a second test, FESST was applied to a time series of FRET efficiencies generated from the MD trajectory. An accurate identification of the native basin of Beta3s is possible with FRET efficiencies calculated from about 1000 photons emitted during the folding time. The analysis of a simple two-state model indicates that the required photon rate is sensitive to the separation of the basins' minima. Therefore, the overlap of the FRET efficiency distributions in the different basins of Beta3s might be responsible for the large number of photons required. In the simple model, a minor increase of the separation of the basins' minima reduces the required photon rate to 200 photons per folding time (the potentials' shape contributes only weakly to the requirements on the photon rate). Extrapolated to current single-molecule FRET experiments, we expect a pronounced discriminatory power of FESST for molecular systems with dynamics in the millisecond time range and above.

The present analysis focused on the FRET efficiency, because it is one of the most commonly used observables. Additional information, e.g., inter-photon times, polarization or fluorescence lifetimes, are expected to further increase the discriminatory power of FESST. In conclusion, FESST can be applied to the time series of any type of scalar observable as long as the short-time distribution of the single-

molecule signal contains enough information to allow FESST to remove the signal's degeneracy.

Acknowledgments

We thank Drs. I. Gopich, S. Muff, D. Nettels and S. V. Krivov for interesting discussions. This work was supported by grants of the Swiss National Science Foundation to A.C. and B.S., and a Starting Investigator Grant of the European Research Council (FP7) to B.S. Most of the simulations were carried out on the Matterhorn computer cluster of the University of Zurich.

References

[1] Goldstein M (1969) Viscous Liquids and the Glass Transition: A Potential Energy Barrier Picture. *The Journal of Chemical Physics* 51:3728–3739.

[2] Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254:1598–1603.

[3] Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4:10–19.

[4] Shakhnovich EI (1994) Proteins with selected sequences fold into unique native conformation. *Phys Rev Lett* 72:3907–3910.

[5] Krivov SV, Karplus M (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc Natl Acad Sci USA* 101:14766–14770.

[6] Rao F, Caflisch A (2004) The protein folding network. *J Mol Biol* 342:299–306.

[7] Krivov SV, Karplus M (2006) One-dimensional free-energy profiles of complex systems:Progress variables that preserve the barriers. *J Phys Chem B* 110:12689–12698.

[8] Caflisch A (2006) Network and graph analyses of folding free energy surfaces. *Curr Opin Struct Biol* 16:71–78.

[9] Bai C, Wang C, Xie XS, Wolynes PG (1999) Single molecule physics and chemistry. *Proc Natl Acad Sci USA* 96:11075–11076.

[10] Ha T, Enderle T, Ogletree DF, Chemla DS, Selvin PR, Weiss S (1996) Probing the interaction between two single molecules:fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc Natl Acad Sci USA* 93:6264–6268.

[11] Deniz AA, Laurence TA, Beligere GS, Dahan M, Martin AB, Chemla DS, Dawson PE, Schultz PG, Weiss S (2000) Single-molecule protein folding:diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc Natl Acad Sci USA* 97:5179–5184.

[12] Schuler B, Lipman EA, Eaton WA (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature (London)* 419:743–747.

[13] Haran, G (2003) Single-molecule fluorescence spectroscopy of biomolecular folding. *Journal of Physics:Condensed Matter* 15:R1291–R1317.

[14] Michalet X, Weiss S, Jäger M (2006) Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem Rev* 106:1785–1813.

[15] Nettels D, Gopich IV, Hoffmann A, Schuler B (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc Natl Acad Sci USA* 104:2655–2660.

[16] Schuler B, Eaton WA (2008) Protein folding studied by single-molecule FRET. *Curr Opin Struct Biol* 18:16–26.

[17] Baba A, Komatsuzaki T (2007) Construction of effective free energy landscape from single-molecule time series. *Proc Natl Acad Sci USA* 104:19297–19302.

[18] Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc Natl Acad Sci USA* 97:10780–10785.

[19] Muff S, Caflisch A (2008) Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics* 70: 1185–1195.

[20] De Alba E, Santoro J, Rico M, Jiménez MA (1999) De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Science* 8:854–865.

[21] Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* 46: 24–33.

[22] Muff S, Caflisch A (2009) ETNA:Equilibrium transitions network and Arrhenius equation for extracting folding kinetics from REMD simulations. *J Phys Chem B* 113:3218–3226.

[23] Cavalli A, Haberthür U, Paci E, Caflisch A (2003) Fast protein folding on downhill energy landscape. *Protein Science* 12:1801–1803.

[24] Hartigan J (1975) Clustering algorithms. *Wiley, New York*.

[25] Krivov SV, Muff S, Caflisch A, Karplus M (2008) One-Dimensional Barrier Preserving Free-Energy Projections of a beta-sheet Miniprotein: New Insights into the Folding Process. *J Phys Chem B* 112:8701–8714.

[26] Krivov SV, Karplus M (2008) Diffusive reaction dynamics on invariant free energy profiles. *Proc Natl Acad Sci USA* 105:13841–13846.

[27] Brooks BR, et al. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217.

[28] Brooks BR, et al. (2009) CHARMM: The biomolecular simulation program. *J Comput Chem* 30:1545–1614.

[29] Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) Wordom:a program for efficient analysis of molecular dynamics simulations. *Bioinformatics* 23:2625–2627.

[30] Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. *J Chem Phys* 105:1902–1921.

[31] Smirnov NV (1939) On the deviations of the empirical distribution curves. *Matematicheskii Sbornik* 6:3–24.

[32] Vershik A (2006) Kantorovich metric: initial history and little-known applications. *J Math Sci* 133:1410–1417.

[33] Gopich IV, Szabo A (2009) Decoding the Pattern of Photon Colors in Single-Molecule FRET. *J Phys Chem B* 113:10965–10973.

[34] Joo C, et al. (2006) Real-Time Observation of RecA Filament Dynamics with Single Monomer Resolution. *Cell* 126:515–527.

[35] Joo C, Balci H, Ishitsuka Y, Buranachai C, Ha T (2008) Advances in Single-Molecule Fluorescence Methods for Molecular Biology. *Annual Review of Biochemistry* 77:51–76.

[36] Borgia A, Williams PM, Clarke J (2008) Single-Molecule Studies of Protein Folding. *Annual Review of Biochemistry* 77:101–125.

[37] Chung HS, Louis JM, Eaton WA (2009) Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc Natl Acad Sci USA* 106:11837–11844.

[38] Schuler B (2007) Application of single molecule Förster resonance energy transfer to protein folding. *Methods Mol Biol* 350:115–138.

[39] Cavalli A, Ferrara P, Caflisch A (2002) Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins: Structure, Function, and Bioinformatics* 47:305–314.

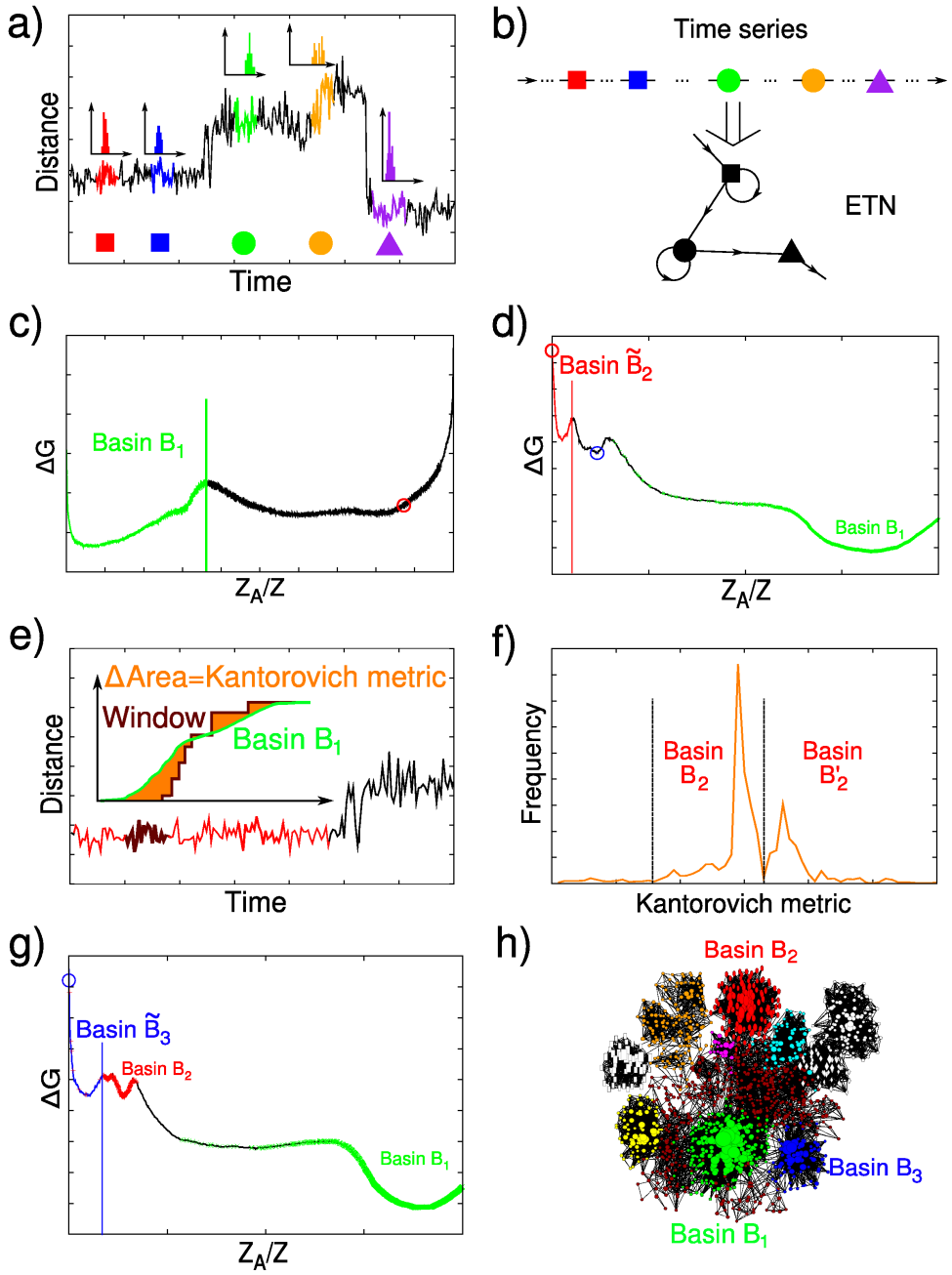

Figure 1: Schematic illustration of the FESST procedure (free energy surfaces from single-molecule time series). (a) The time windows of the scalar signal are coarse-grained according to the short-time distribution of the distance. (b) The coarse-graining yields a time series of nodes and transitions, which define the ETN. (c) The cFEP is plotted using the most populated node as a reference. The first free energy basin is isolated by cutting at the first barrier. The red circle indicates the most populated node outside the first basin, which is used to plot the cFEP for the determination of the second basin. (d) Due to the degeneracy of the short-term distribution, nodes from different free-energy basins overlap in the second basin (see text). The tilde is used to denote a cFEP basin with overlap. The blue circle is the most populated node outside of $\tilde{B}_2$. (e,f) The overlap in $\tilde{B}_2$ is removed by comparing with the whole distribution of the first basin (B_1). (g) The procedure is repeated for the next basin. (h) The basins extracted by FESST are illustrated on the conformation space network of Beta3s with the native basin in green, and non-native basins Ch-curl$_1$ (B_2) and Ns-or$_1$ (B_3) in red and blue, respectively [19].

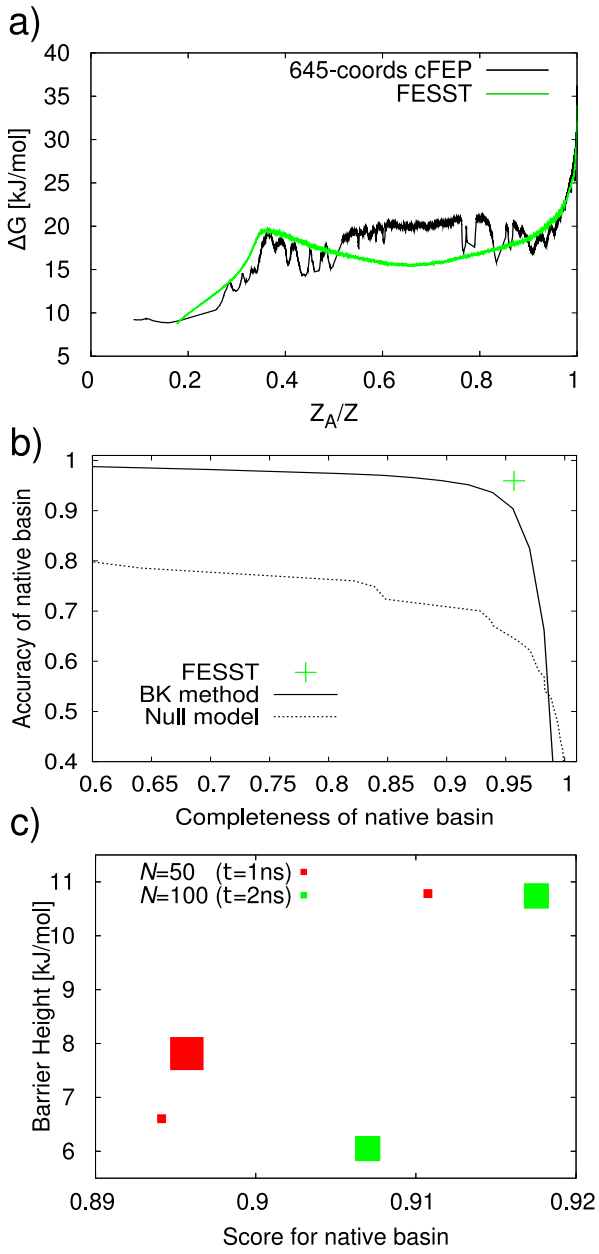

Figure 2: (a) Determination of the native basin by taking into account all structural information (black) or only a single distance (green). The cFEP is shown with the most populated node as a reference, and its determination is presented in the SI. (b) Comparison of FESST with a previously published single distance approach (BK= Baba and Komatsuzaki [17]) and the null model. The null model reports the accuracy and completeness if the native basin is selected according to a range of distances from a comparison of histograms after projection on a single coordinate (Fig. 3a). Note that FESST yields a single data point rather than an accuracy vs. completeness curve, because the native basin is defined by the unique location of the first peak in the cFEP. (c) Effect of the parameters used for coarse-graining on the FESST determination of the native state and the height of the unfolding barrier on the cFEP. Note that multiple values of the threshold ζ yield the same score and barrier height. The size of the symbol is proportional to the number of values tested. As an example, the best result, i.e., the data point with highest score *and* barrier height (green square in the top right corner) is obtained with $\zeta = 0.30, \zeta = 0.32,$ and$\zeta = 0.34$ using a window size of 100. The plot provides evidence that the parameters can be optimized with the height of the cFEP barrier as a cost function.

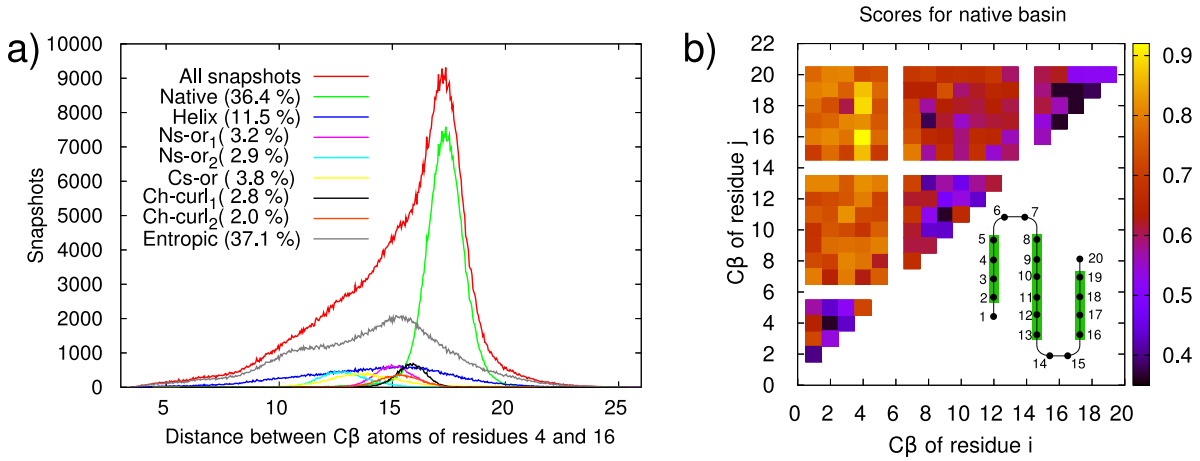

Figure 3: Robustness of FESST with respect to the choice of the distance. (a) Histograms of distance between the C_β atoms of residues 4 and 16 for the snapshots in each free-energy basin determined by cFEP using all 645 degrees of freedom of Beta3s [25]. (b) Matrix of scores for native state detection. Each (i,j) value of the score was calculated by applying FESST to the time series of distance between the C_β-atoms of residues i and j (Gly_6 and Gly_{14} have no C_β atom). The inset shows a schematic representation of Beta3s with the three native β-strands (green rectangles). In the coarse-graining step of FESST, N = 100 distance values are compared and the acceptance cutoff $\zeta = 0.3$ is used in the Kolmogorov-Smirnov test (see Methodology).

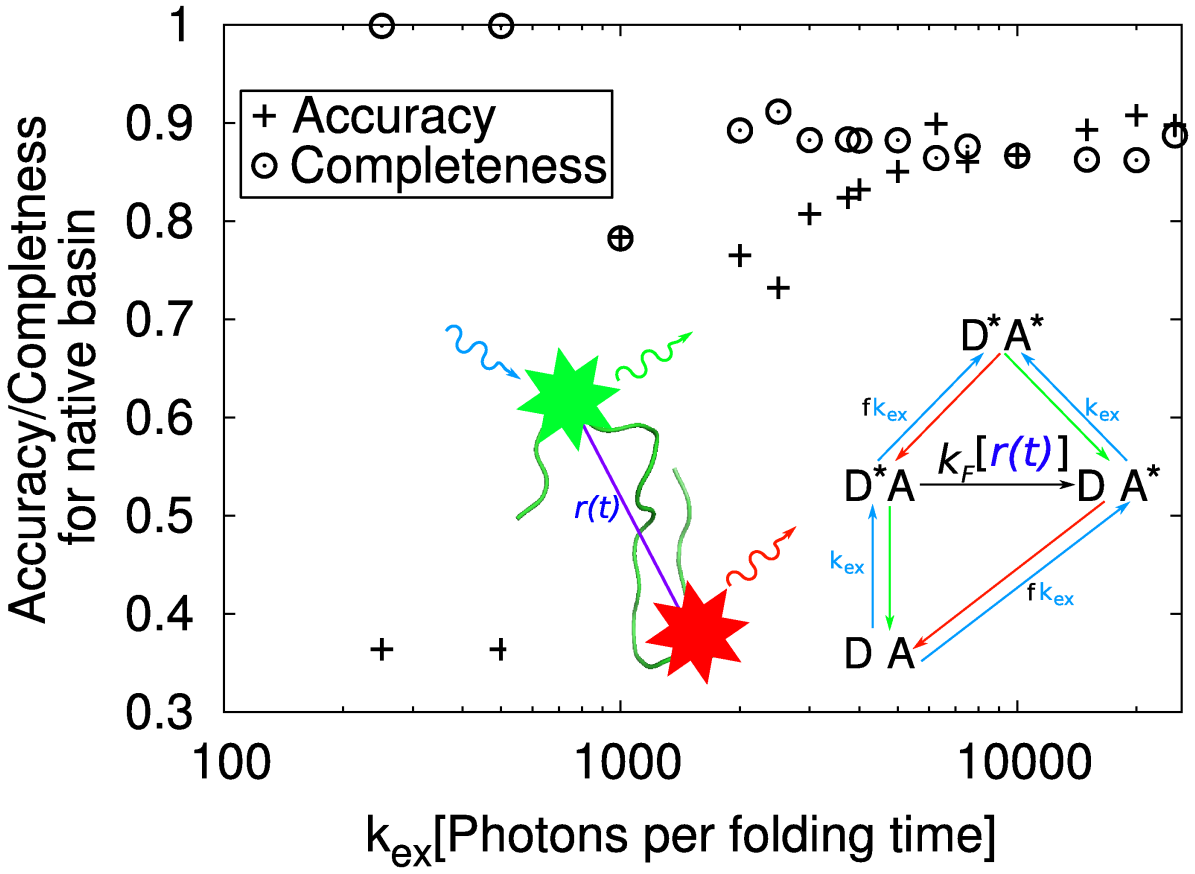

Figure 4: FESST performance on a simulated FRET experiment. The time series of FRET efficiencies calculated for 0.4-ns bins is used for the analysis by FESST with a window size of 25 bins (the effect of other window sizes is illustrated in Fig. S5) and acceptance cutoff $\zeta = 0.3$. The accuracy and completeness for the identification of the native basin is calculated for the set of snapshots in all bins of the first FESST basin. The excitation rate k_{ex} is expressed as the average number of photons per folding time (about 100 nanoseconds for Beta3s [19] in MD simulations at 330 K). The left inset illustrates the distance r between the C_β atoms of residues 4 and 16 determined from an MD simulation. The Markov state model of the photophysics depicted as inset on the right is used to simulate a FRET experiment and obtain photon arrival times. The Förster rate $k_F[r(t)]$ is the distance-dependent rate of the energy transfer from the donor to the acceptor chromophore (see SI for a detailed description). It is important to note that each excitation leads to an emitted photon in the FRET emulation. Direct excitation of the acceptor is set to f = 5% of the donor excitation rate [38].

Supporting Information

Molecular Dynamics simulations

Ten MD runs of 2 μs each with different initial distributions of velocities were performed with the Berendsen thermostat (coupling constant of 5 ps) at 330K, which is slightly above the melting temperature of Beta3s [39]. A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 10^6 MD snapshots. This required three weeks on a 10-CPU cluster.

Merging of native nodes

In comparison to the 645-coordinates ETN, the FESST-ETN lacks nodes with very high weight (Fig. S6). For instance, the most populated node in the FESST-ETN (158 snapshots) is around 557 times smaller than most visited node of the 645-coordinate ETN (88022 snapshots). Therefore, the free energy basin represented in the cFEP is too shallow and the height of the unfolding barrier is underestimated (inset in Fig. S2). To render the most populated node in the FESST-ETN more representative of the native basin, the M heaviest native nodes are combined. Per definition, native are those nodes with a value of the progress variable Z_A/Z in the cFEP (calculated from the most populated node) smaller than the value at the first peak. The new transition network is constructed from the node sequence in the MD simulation with the heaviest M native nodes merged to one node.

This merging step affects predominately the native basin (inset of Fig. S2). The largest value for the height of the unfolding barrier is observed for M = 7000 nodes merged. The value of the barrier height is robust for $5000 \leq \mathrm{M} \leq 10000$. For $\mathrm{M} \geq 3000$ nodes merged, the weight of the most populated node in the FESST-ETN exceeds the weight of the most visited node in the unmodified 645-coordinates ETN (Fig. S7). The merging procedure can also be applied to the 645-coordinates ETN. The highest barrier is found for 107 nodes merged and exceeds the value for the unmodified network by only 0.7 kJ/mol (Fig. S2).

A second consequence of the split reference node is the significant increase of

the time to reach the reference node from any node in the time series of states (Fig. S8). For the ETNs with M $\geq$ 3000, the distribution of the mean first passage times matches those of the 645-coordinates ETN (Fig. S8). The correspondence of the folding time distributions fortifies the assumption that the system dynamics is reliably captured by the FESST coarse-graining. The reliable representation of the system's dynamics in the ETN is a necessary condition for the correct operation of the cFEP approach [7].

Computational costs

Coarse-graining is the computational bottleneck, and the time it requires depends on the parameters used. Coarse-graining of the time series (one million time windows) of the distance between C_β-atoms of residues 4 and 16 with a window size of N = 100 and a cutoff parameter $\zeta = 0.3$ takes 6 hours on a recent XEON CPU with 2.33 GHz clock frequency. Elevating the cutoff to $\zeta = 0.38$ reduces the running time to 4.5 hours. The cFEP calculation takes only five to ten minutes. Very small memory requirements are needed for both procedures. Note that the determination of multiple free energy basins requires only one coarse-graining, but multiple cFEP calculations.

Simulating FRET experiments

The photon arrival times are generated by a random walker on the Markov state model shown in detail in Fig. S9. The states are DA (both donor and acceptor in ground state), D^*A (donor in excited state, acceptor in ground state), DA^* and D^*A^*. The transition probabilities are approximated by the product of the transition rate and the time step (chosen to be $dt = 0.2$ ps, i.e., 100 observations along the MD saving interval of 20 ps). Finer time steps changed the photon counting results only marginally. For the relaxation rates of donor and acceptor, $k_A = k_D = 2500\frac{1}{2\,\mathrm{ns}}$ is used. Direct excitation of the acceptor is set to five percent of the donor excitation rate. The Förster rate $k_F(r) = k_D \left(\frac{R_0}{r}\right)^6$ is calculated from the instantaneous distance r between the C_β-atoms of residues 4 and 16, and $R_0 = 12$ Å

is the Förster radius. This Förster radius $R_0 = 12$ Å is the value most remote from the native peak that separates the distributions of FRET efficiencies of native and non-native conformations best. This separation is important, because there is an anticorrelation of the score of the native basin and overlap of the distributions in native and non-native state (data not shown).

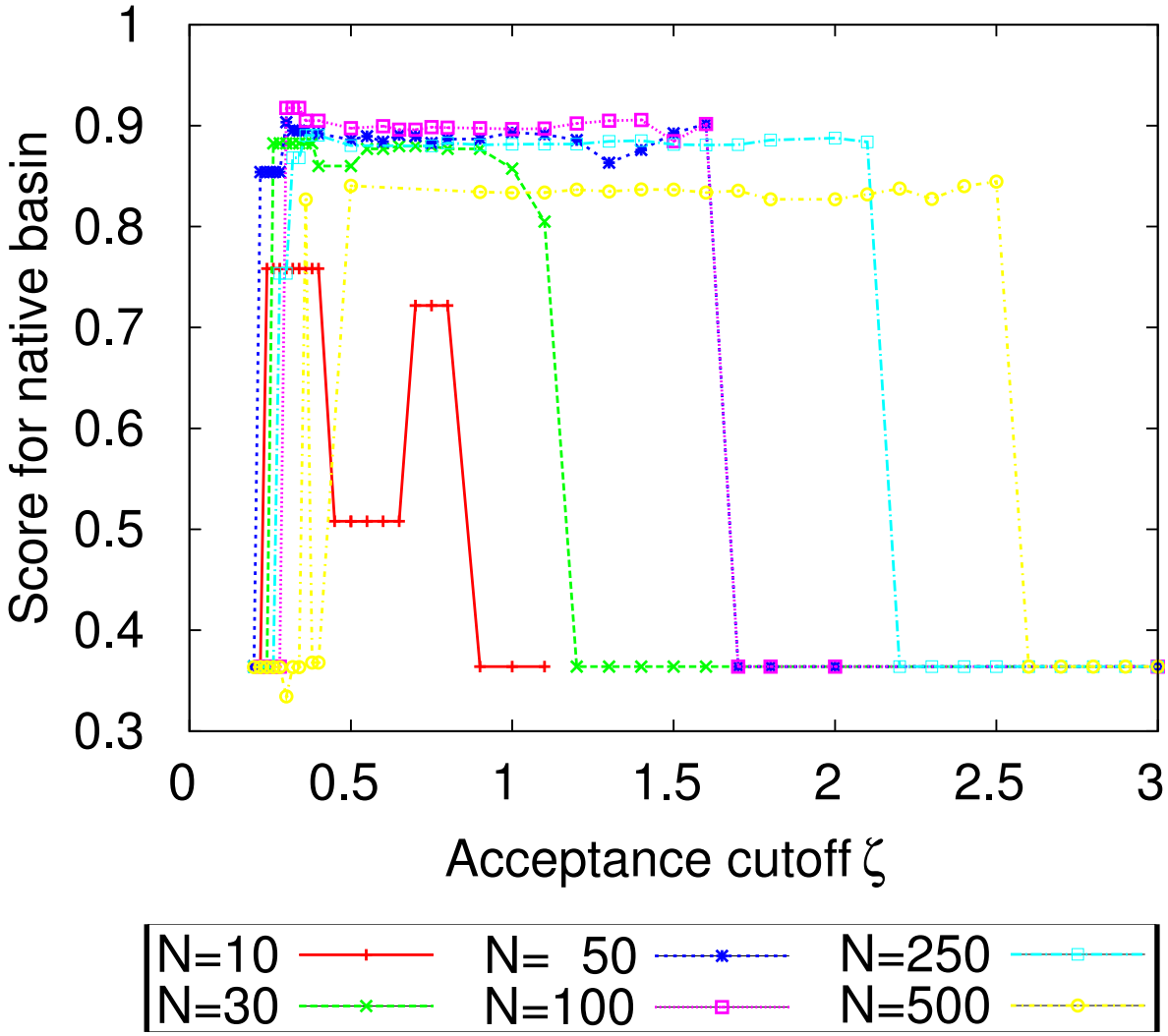

Figure S1: Robustness of FESST upon variation of the parameter used for coarse-graining. The range of values tested for the size of the time window is $10 \leq \mathrm{N} \leq 500$, i.e., 0.2 ns $\leq \tau \leq$ 10 ns as the number of MD snapshots N is equal to τ times the saving frequency of 1/20 ps^{-1}. Values of $\tau = 2$ns (N=100) and $\zeta = 0.3$ are used in most of the main text.

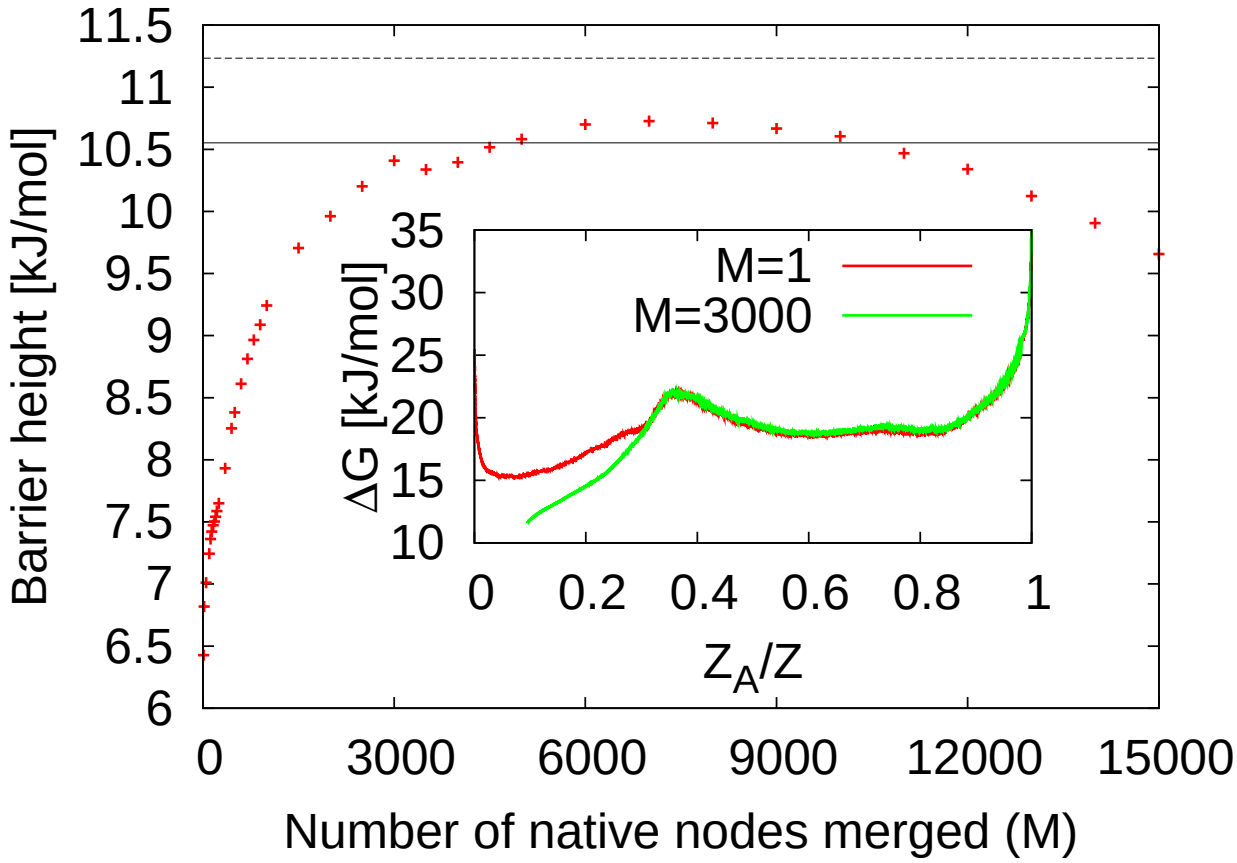

Figure S2: Dependence of barrier height on the merging of the heaviest nodes of the native basin. The solid horizontal line indicates the barrier height of the 645-coords cFEP [25]. The dashed horizontal line displays the maximal barrier height found, which results when the heaviest 107 native nodes in the 645-coords cFEP are merged. The inset shows the FESST cFEPs with M= 3000 native nodes merged (green) and without merging (red).

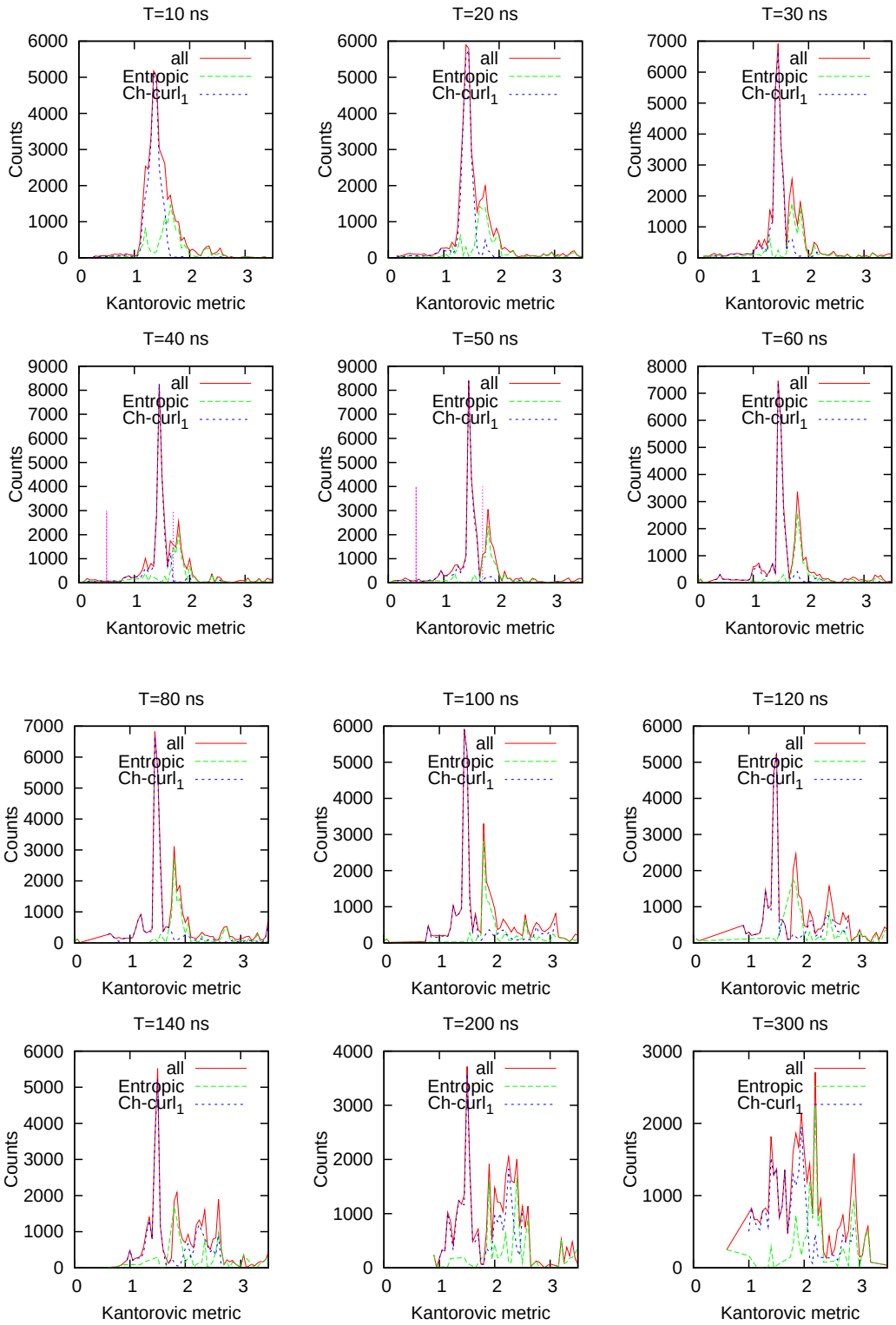

Figure S3: Effect of different window lengths in basin overlap removal. Compared is the Kantorovich metric distribution (area between cumulative histograms) between the long-time distance distributions in windows of varying length T around MD snapshots in $\tilde{\mathrm{B}}_2$ and the native distance distribution, i.e. all MD snapshots in B_1 (see Fig. 1 for definition of $\mathrm{B}_1, \tilde{\mathrm{B}}_2$). This plot illustrates that the ranges of Kantorovich metric are different for the different subbasins in $\tilde{\mathrm{B}}_2$, Ch-curl$_1$ and Entropic.

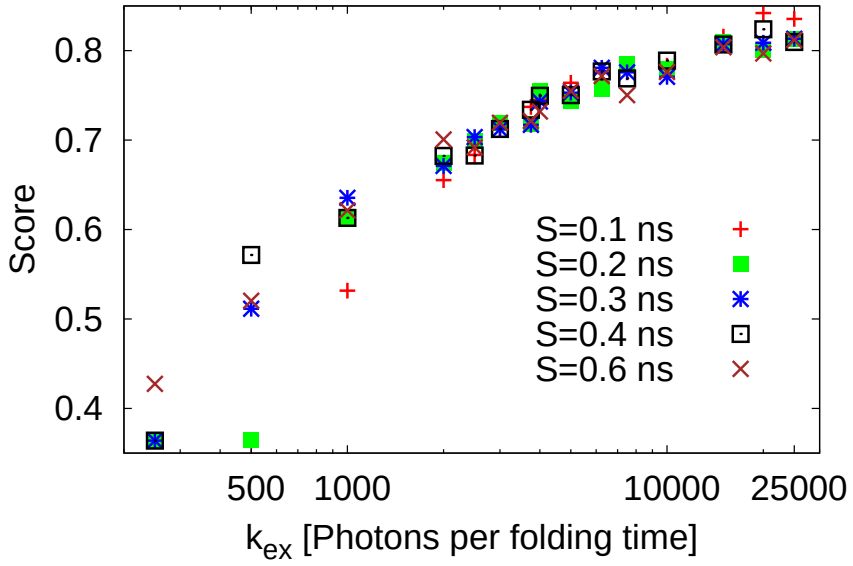

Figure S4: Effect of different binning times S on the score of the native basin detection for simulated FRET experiments. Note that the scores are calculated on a snapshot-wise basin assignment as for Fig. 4. For each excitation rate k_{ex}, only the highest score (tested are window sizes of $2, 4, 6, 8, 10, 20$, and 40 ns) is shown. The excitation rate k_{ex} is expressed as the number of photons emitted per folding time, which is 100ns for Beta3s at 330K (Fig. 4).

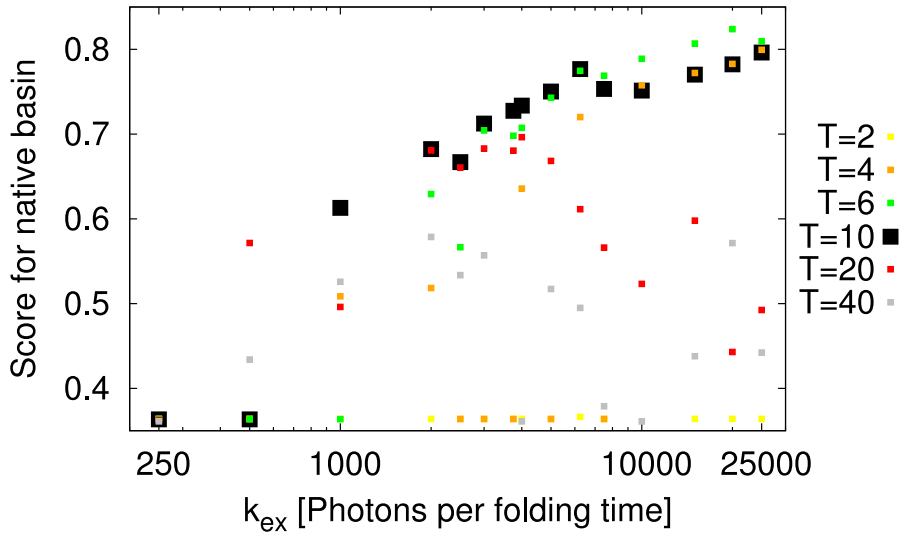

Figure S5: Effect of different window sizes T [ns] on FESST performance in emulated FRET experiments with 0.4 ns bins. The identification of the native basin is robust with respect to the choice of the window size in the range 4 ns $\leq$ T $\leq$ 10 ns. The same setup as for the data shown in Fig. 4 and described in the SI is used. As in Fig. 4, the excitation rate k_{ex} is expressed as the number of photons emitted during the folding time, which is 100ns for Beta3s at 330K.

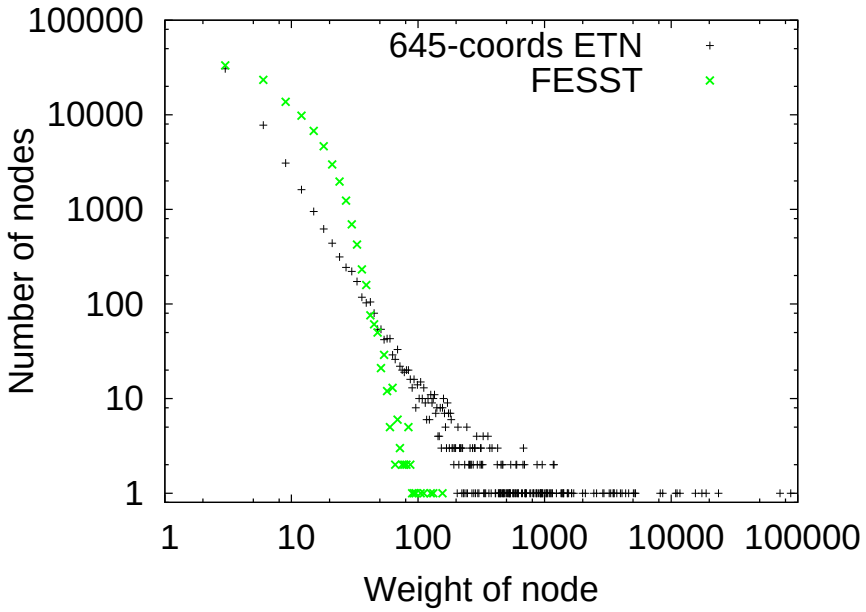

Figure S6: Distribution of node weights (number of snapshots) for the 645-coords ETN and the FESST ETN (distance 4.$C\beta$-16.$C\beta$ in Beta3s, window size $N = 100$, and acceptance cutoff $\zeta = 0.3$).

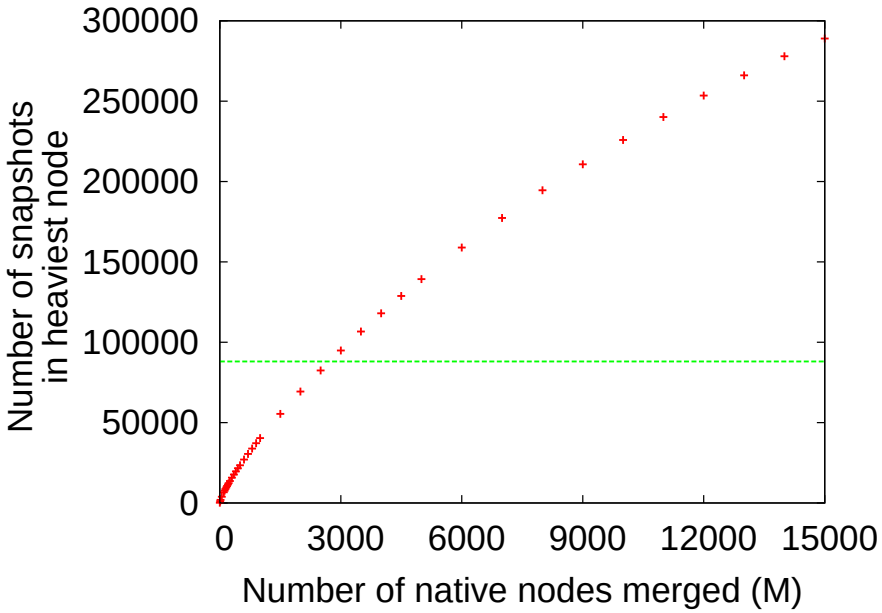

Figure S7: Dependence on the number of merged native nodes M of the number of snapshots in the heaviest node of the FESST-ETN, extracted from the time series of the distance between the C_β atoms of the residues 4 and 16 in Beta3s with window size $N = 100$ and acceptance cutoff $\zeta = 0.3$. The horizontal line indicates the number of snapshots in the heaviest node of the 645-coordinates ETN.

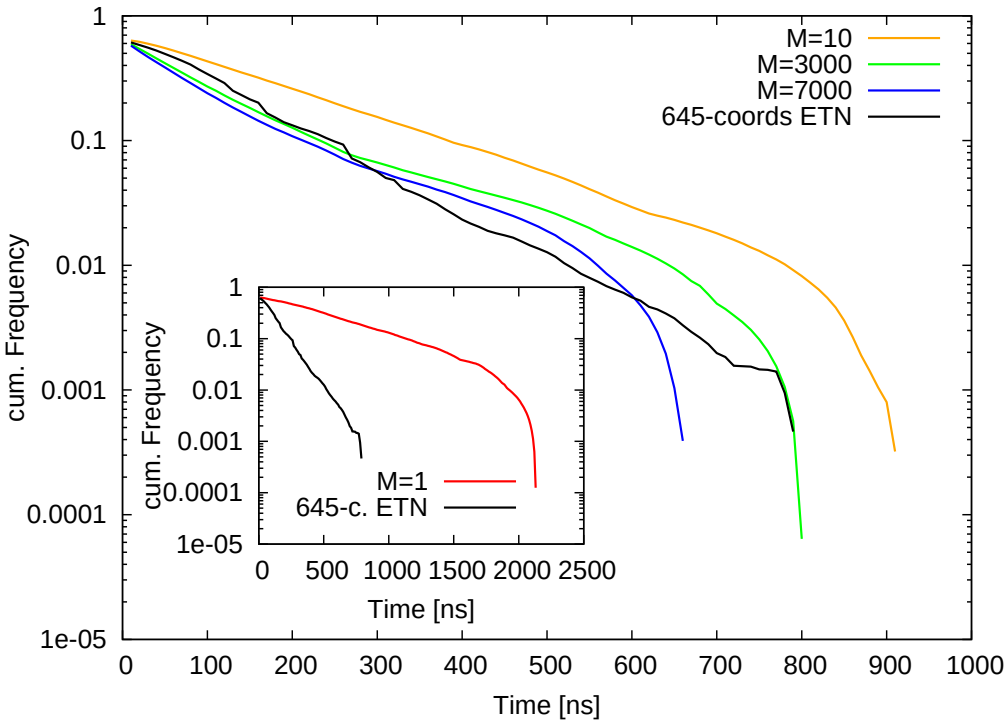

Figure S8: Cumulative distribution of first passage times to the native node for the 645-coordinates ETN [25], and the FESST-ETNs with M= 10, 3000, and 7000 native nodes merged. The inset shows the much slower folding of the FESST-ETN without native node merging (M=1), which is a consequence of the small weight of the most populated nodes (Fig. S7).

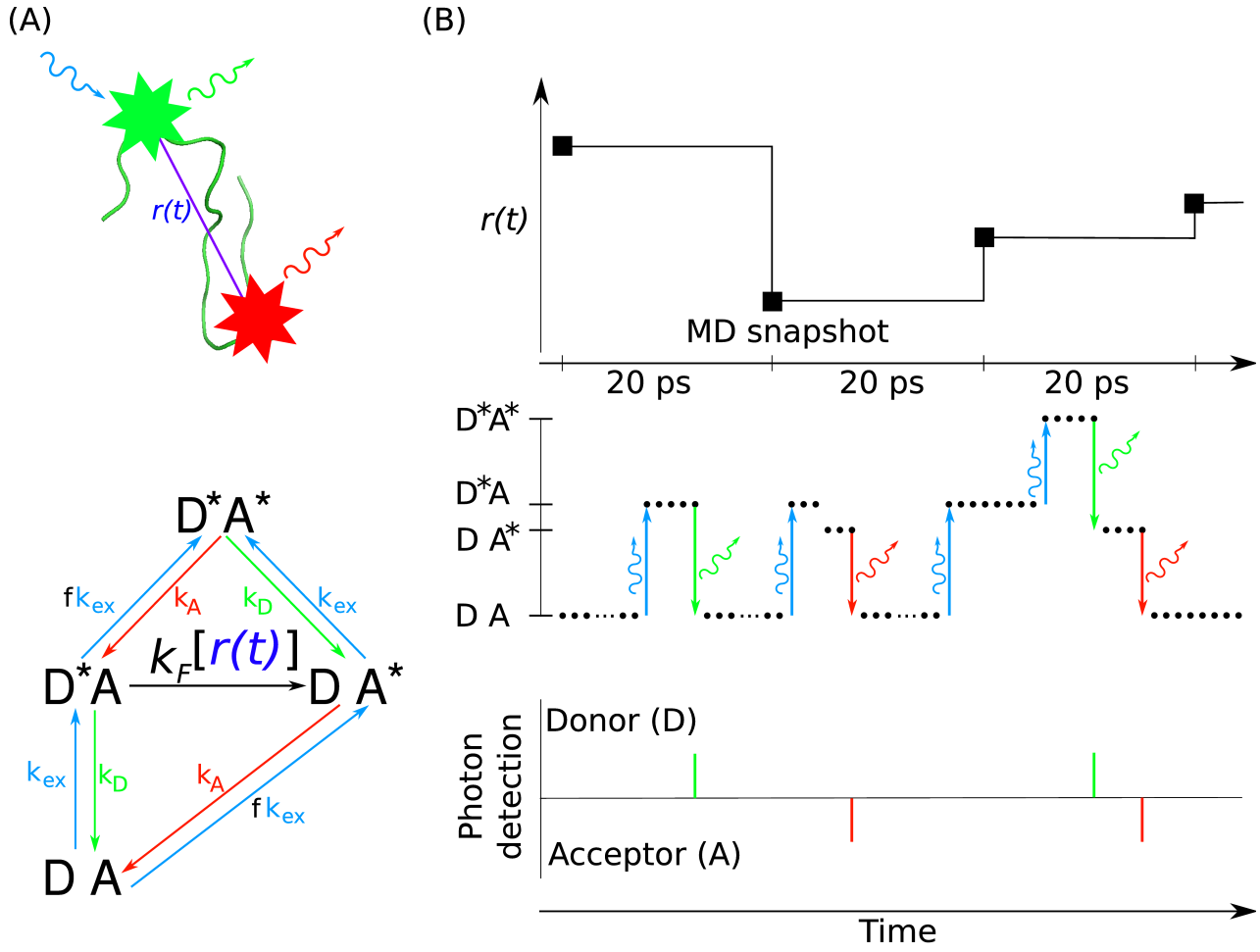

Figure S9: Markov state model used to generate the photon time series. Red, green, and blue curled arrows represent acceptor photon emission, donor photon emission, and photon absorbance. (A,top) Schematic illustration of the emulated FRET experiment, where $r(t)$ represents the distance between the C_β-atoms of residues 4 and 16. (A,bottom) State diagram of the Markov process used to simulate the FRET experiment. (B) Illustration of the simulation of the FRET experiment, which uses the time series of the distance r, measured along the MD trajectory, to generate the photon time series. For each MD saving interval of 20 ps, 100 steps (circles) of a random walker on the Markov state model are carried out with a constant value of the distance $r(t)$, i.e., constant $k_F[r(t)]$.

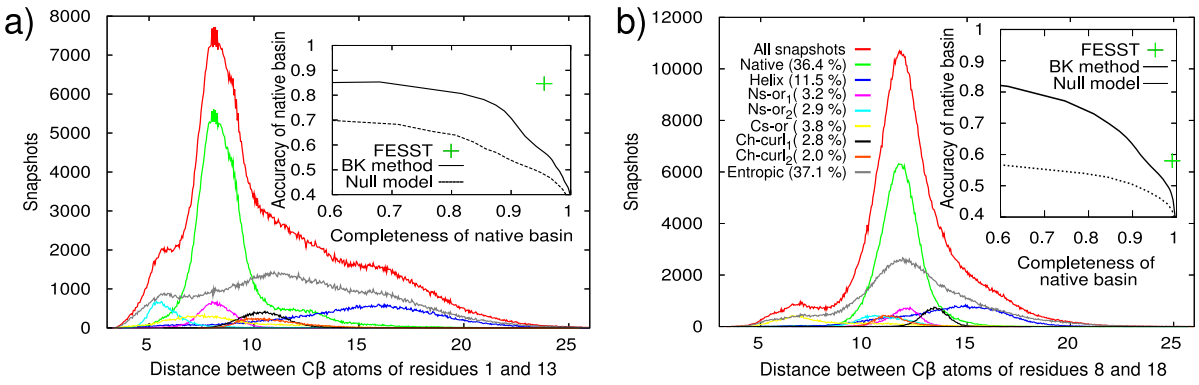

Figure S10: Distribution of inter-residue distances in different free-energy basins as identified using all 645 coordinates of Beta3s. As inset the accuracy and completeness of the native state detection of FESST is compared with the BK procedure [17] and a null model.

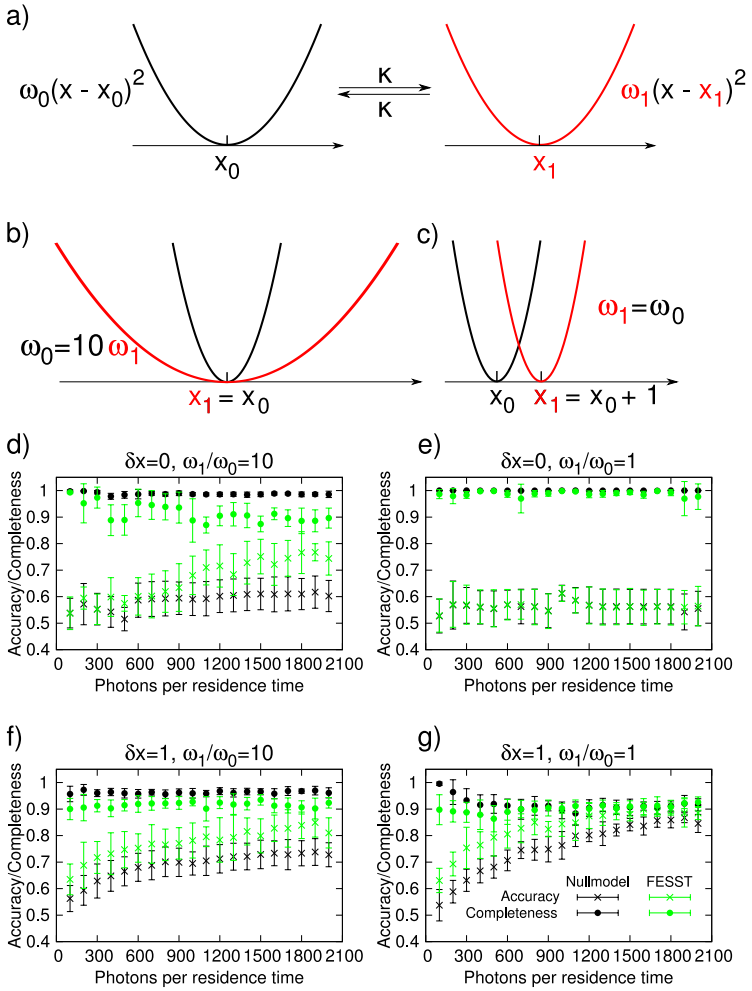

Figure S11: Resolution limits of FESST examined with a simple two-state model. (a) Schematic illustration of the model: Dynamics of a massive particle in a harmonic potential that switches from one shape to the other with rate κ. The position time series is then transformed to a time series of FRET efficiencies as for Beta3s (the equilibrium position is defined as $x_0 = 10$ A.U., the Förster radius as $R_0 = 12$ A.U., the curvature of potential 0 as $\omega_0 = 100$ A.U. and the interchange rate is assumed to be $\kappa = 0.005$ A.U.). (b) Illustration of the "perfect" overlap situation, where the equilibrium positions in both potentials are equal and the curvatures differ by a factor of ten. (c) Illustration of the effect of a shift in the equilibrium position of the potentials. (d-g) Accuracy and completeness of the basin with index 0 as detected by FESST and the best null model in ten independent simulations. The parameters for FESST are optimized using the height of the unfolding barrier determined intrinsically (cf. Fig. 1.c). For the null model, the basin is formed by all bins with a FRET efficiency lower than a given cutoff. To make the most stringent comparison, the detection quality of the null model is maximised by finetuning both the length of the binning interval and the cutoff value based on the knowledge of the solution. For a minor shift $\delta x = 1$ of the two minima of the potential (panels (f) and (g)), FESST yields significantly more accurate solutions with only slightly lower completeness than the null model for as few as 200 photons per residence time. An increased amount of about 1000 photons per residence time is required if the equilibrium positions match and the curvatures differ by a factor of ten (panel d). For identical potentials, the solutions of FESST and the null model are equal (panel e).

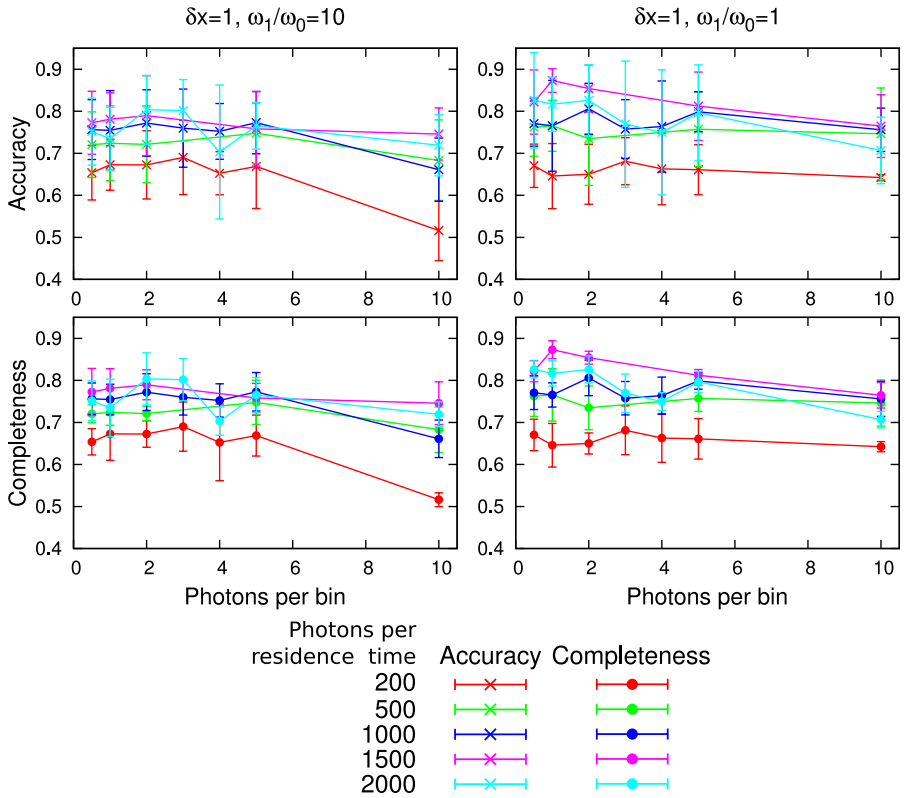

Figure S12: Dependence of FESST performance in the simple two-state model (cf. Fig. S11) on the number of photons per FRET bin. The accuracy and completeness of the basin with index 0 as detected by FESST depend mainly on the number of photons emitted during the residence time and only weakly on the explicit number of photons per bin.

Chapter 3

Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement.

P. Schuetz and A. Caflisch

[*Phys. Rev. E*, **2008**, *77*, 046112]

Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement

Philipp Schuetz and Amedeo Caflisch
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
(Received 7 December 2007; revised manuscript received 22 February 2008; published 17 April 2008)

Identifying strongly connected substructures in large networks provides insight into their coarse-grained organization. Several approaches based on the optimization of a quality function, e.g., the modularity, have been proposed. We present here a multistep extension of the greedy algorithm (MSG) that allows the merging of more than one pair of communities at each iteration step. The essential idea is to prevent the premature condensation into few large communities. Upon convergence of the MSG a simple refinement procedure called "vertex mover" (VM) is used for reassigning vertices to neighboring communities to improve the final modularity value. With an appropriate choice of the step width, the combined MSG-VM algorithm is able to find solutions of higher modularity than those reported previously. The multistep extension does not alter the scaling of computational cost of the greedy algorithm.

DOI: 10.1103/PhysRevE.77.046112 PACS number(s): 89.75.Fb, 05.10.−a, 89.75.Hc

I. INTRODUCTION

The networks under study in natural and social sciences often show a natural divisibility into smaller modules (or communities) originating from an inherent, coarse-grained structure. In general, these modules are characterized by an abundance of edges connecting the vertices within individual communities in comparison to the number of edges linking the modules.

To detect these partitions several algorithm- or score-based approaches have been developed and applied. Very popular became the approach introduced by Girvan and Newman [1] based on the quality function called "modularity" for partition assessment. This scoring function compares the actual fraction of intracommunity edges with its expectation in the random case given an identical degree distribution. The partition with the highest value of the scoring function is then considered to be the optimal splitting. The modularity Q is defined (for undirected networks) as

$$Q = \sum_{i=1}^{N_C} \left[\frac{I(i)}{L} - \left(\frac{d_i}{2L} \right)^2 \right]$$

with $I(i)$ the weights of all edges linking pairs of vertices in community i, d_i the sum over all degrees of vertices in module i, L the total weight of all edges, and N_C the number of communities.

Intrinsically, the modularity based approach does not prescribe the usage of a particular optimization procedure. In practice, a strategy for optimization has to be chosen. The modularity optimization is a NP-hard problem [2]. Therefore, only an exhaustive search reveals the optimal solution for a generic network. This type of search is extremely demanding and only in a few cases feasible. Thus, many heuristic approaches such as extremal optimization [3], simulated annealing [4], and the greedy algorithm [5] have been developed, refined, and successfully applied. Among the published approaches the greedy algorithm is one of the fastest techniques [6]. On the other hand, many examples show that the greedy algorithm is not capable of finding the solutions with the highest modularity value. Furthermore, recent studies have provided evidence that modularity [7] and Potts model based approaches [8] are endowed with an intrinsic resolution limit (small modules are not detected and amalgamated into bigger ones). Thus, each community has to be refined by subduing it as a separate network to the community detection algorithm. Therefore, a fast and accurate optimization technique is necessary.

In this article, we enhance the greedy algorithm by a multistep feature in combination with a local refinement procedure. The enhanced algorithm finds partitions with higher modularity values than previously reported. This paper is organized as follows. In Sec. II we introduce both procedures and describe the motivation for their construction. In addition, we discuss performance oriented implementations and estimate their running times. Benchmarking results for a set of real-world networks and a comparison with other published results are presented in Sec. III. The conclusions are in Sec. IV. In this paper, all networks are considered as undirected. The extension to directed networks is straightforward.

II. THE ALGORITHM

A. Multistep Greedy algorithm (MSG)

Each vertex is a community
Calculate the modularity change matrix ΔQ
Determine the community degrees d_i
while pair (i,j) with $\Delta Q_{ij} > 0$ exists **do**
  for all elements $(i,j,\Delta Q_{ij})$ in ΔQ matrix, parsed with respect to decreasing ΔQ and increasing (i,j)
  do
    if
      $\left\{ \begin{matrix} \Delta Q_{ij} > 0 \text{ in best } l \text{ values in } \Delta Q \text{ matrix} \\ i \text{ and } j \text{ unchanged in iteration} \end{matrix} \right\}$
    then
      MergeCommunities(i,j)
    end if
  end for
end while

Algorithm 1: Flowchart of the MSG algorithm. The modularity change is calculated according to Eq. (1). Details of the algorithm are given in algorithm 2.

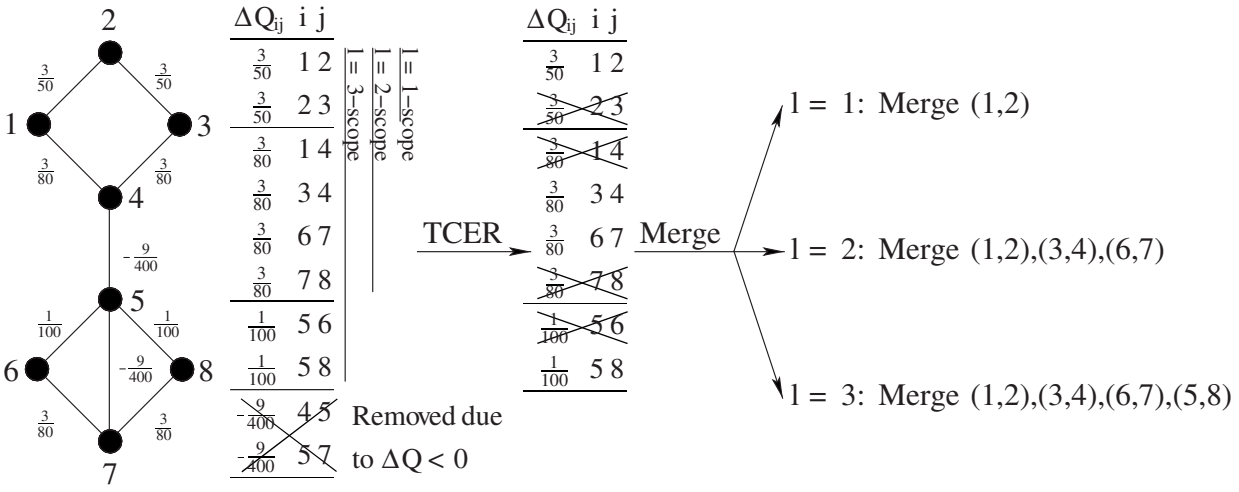

FIG. 1. Effect of different values of level parameter during first MSG iteration on example network.

The classical greedy algorithm (first application in Ref. [5]) joins iteratively the pair of communities that improves modularity most in each step. The essential idea of the "multistep greedy" (MSG) algorithm is to promote the simultaneous merging of several pairs of communities at each iteration. The pseudocode of the MSG algorithm is presented in algorithms 1 and 2, and an illustrative example is given in Fig. 1. The MSG algorithm starts with each vertex separated in its own community. At each iteration the modularity change ΔQ_{ij} upon merge of each pair of connected communities (i,j) is calculated (while nonconnected pairs are ignored because their merging yields a negative modularity change). The triplets $(i,j,\Delta Q_{ij})$ are parsed in the order of decreasing ΔQ value and increasing community index. Those community pairs (i,j) are joined which fulfill the following two criteria:

(1) The modularity change ΔQ_{ij} is within the l most favorable values (levels) and positive.

(2) Touched-community-exclusion-rule (TCER): Neither module i nor j is present in another pair inducing a higher modularity change.

Convergence is reached when all pairwise merges of communities decrease modularity (by induction one can prove that all merges in further iterations would decrease modularity). A level encompasses all triplets $(i,j,\Delta Q_{ij})$ with equal ΔQ_{ij} value and the level parameter l is kept constant. By construction the level parameter is always smaller than the number of edges in the network.

The multiple levels promote the concurrent formation of multiple centers. Simultaneously growing community centers hinder the condensation into few large communities (few formed communities scrape all vertices as the establishment of a new community is too expensive in modularity) as observed in the classical greedy algorithm. The TCER is a second mean against excessive aggregation into few large modules. This rule permits the addition of only one community to an existing community per algorithm iteration. Furthermore, the TCER guarantees that the modularity change upon all performed merges is just the sum over the corresponding ΔQ elements which improves efficiency.

B. Implementation details of MSG

The key observation for an efficient implementation of the MSG is the following: Upon merge of communities i and j only those ΔQ elements concerning either of the two modules have to be recalculated. When the modules i and j are joined into a new one called I, the updated modularity changes $\Delta Q_{Ik}^{\text{new}}$ (module k is connected either to community i or j) reads (see Sec. II in Ref. [9] for details)

$$\Delta Q_{Ik}^{\text{new}} = \begin{cases} \Delta Q_{ik} + \Delta Q_{jk}, & i,j, \text{ and } k \text{ pairwise connected}, \\ \Delta Q_{ik} - \dfrac{d_j d_k}{2L^2}, & i \text{ and } k \text{ connected}, \ j \text{ and } k \text{ not}, \\ \Delta Q_{jk} - \dfrac{d_i d_k}{2L^2}, & j \text{ and } k \text{ connected}, \ i \text{ and } k \text{ not}, \end{cases} \tag{1}$$

with d_x the sum over all degrees of vertices in community $x=i,j$ and L the total edge weight.

Further efficiency improvements are gained from an appropriate choice of data structures. A set (implementation taken from the C++-STL library) is a sorted binary search tree. In a set individual elements can be found or inserted in $O(\log(n))$ time (n the number of elements) and the extremal entries are found in constant time. The modularity changes are stored in the ΔQ matrix implemented as vector of row structures. The ith row consists of a set with elements $(j,\Delta Q_{ij})$ (j a module linked to the community i) ordered according to the community index j. This data structure obsoletes a separate storage of the topology information. The extraction of the best l modularity changes is handled via the *level set*. For each pair of connected communities i and j the element $(\min\{i,j\},\max\{i,j\},\Delta Q_{ij})$ is added to the *level set*. The *level-set* elements are sorted with respect to decreasing ΔQ and increasing index values. The degree information is stored in a vector henceforth named d. In each iteration a

Boolean vector called "touched" stores whether a community has already been modified in the same round. To save the time to determine the highest index of a present communities, the number of vertices (initial length) is chosen as length of the *touched* vector.

The implementation details of the MSG algorithm are listed in algorithm 2. The calculation of the community degrees involves one parse of the edge information. In the second parse of the edge information the ΔQ matrix and the *level set* is filled. The initial modularity change ΔQ_{ij} upon join of modules (at this stage the vertices) i and j is calculated as (see Sec. II in Ref. [9] for details)

$$\Delta Q_{ij} = \frac{I}{L} - \frac{d_i d_j}{2L^2}$$

with I the weight of the edges connecting the vertices i and j, d_x the degree of vertex $x=i,j$, and L the total edge weight. The modularity value of the initial partition is (N the number of vertices)

$$Q_0 = -\sum_{i=1}^{N} \frac{d_i^2}{4L^2}.$$

Each vertex is a community
Calculate community degrees d and the ΔQ matrix
Determine the initial modularity $Q \leftarrow Q_0 = -\Sigma_{i=1}^{n} \frac{d_i^2}{4L^2}$
level set $\leftarrow$ set of ΔQ elements $(i,j,\Delta Q_{ij})$, sorted with respect to decreasing ΔQ and increasing (i,j)
while first element of *level set* has $\Delta Q > 0$ **do**
  $touched \leftarrow (0,\ldots,0)$ Boolean,
  N-dimensional vector (N= No. vertices)
  $\{touched_i = 1$, if module i is modified in *while*-loop$\}$
  $MP \leftarrow$ subset of *level-set* elements $(i,j,\Delta Q_{ij})$ with $\Delta Q_{ij} > 0$ and ΔQ_{ij} among highest l values
  for all elements $(i,j,\Delta Q_{ij})$ of MP **do**
    if (**not** $touched_i$) **and** (**not** $touched_j$) **then**
      while parse $\Delta Q_{i.}$ and $\Delta Q_{j.}$ concurrently **do**
        ΔQ_{ik}

$$\leftarrow \begin{cases} \Delta Q_{ik} + \Delta Q_{jk}, & i,k \text{ and } j,k \text{ are linked,} \\ \Delta Q_{ik} - \frac{d_j d_k}{2L^2}, & i \text{ and } k \text{ are linked,} \\ \Delta Q_{jk} - \frac{d_i d_k}{2L^2}, & j \text{ and } k \text{ are linked,} \end{cases}$$

        $\Delta Q_{ki} \leftarrow \Delta Q_{ik}$
        Update the *level set*
        Update the modularity $Q \leftarrow Q + \Delta Q_{ik}$
      end while
      Empty $\Delta Q_{j.}$
      Flag $touched_i, touched_j \leftarrow 1$
      Update degrees: $d_i \leftarrow d_i + d_j, d_j \leftarrow 0$
    end if
  end for
end while

Algorithm 2: Performance-oriented implementation of MSG algorithm. The touched vector contains the information for the touched-community-exclusion-rule (TCER).

The algorithm iteration starts by initializing the *touched* vector. Subsequently, the *level set* is parsed and all elements with positive ΔQ value, whose modularity change is among the best l (external level parameter) different values, are stored in a set named MP conserving the order of the *level set*. In this order the module pairs are merged unless one of them was part of a amalgamation in the same algorithm iteration. In the merge process, the changed ΔQ matrix elements are calculated as described at the beginning of this paragraph. To determine which case applies in Eq. (1) the fact that each row of the ΔQ matrix is ordered with respect to the community index can be used. More precisely, parse for the merge of modules i and j the corresponding rows concurrently. For each row define an momentarily considered element p. If the community index of p_i is equal to the one of p_j, the first case applies and advance both p's to the next element in the corresponding row. If the index k of p_i is lower than the one of p_j calculate the $\Delta Q_{Ik}^{\text{new}}$ element (I the name of the merged community) according to the second case and advance (if possible) only p_i. If the module index of p_i is larger than the one of p_j, proceed analogously. If one p reaches the end of the row, merge the remaining elements of the other row according to the respective rule. This procedure will be called "asynchronous parsing" in Sec. II C. It is customary to update each ΔQ element after calculation. To complete the merge process it remains to update the community degrees and to flag the modified communities in the touched vector.

C. Running time estimation of MSG

As we adopted the modularity change calculation of Clauset *et al.* (Sec. II in Ref. [9]) we can adopt their method of running time estimation as well. First, we observe that the update of one element in the ΔQ matrix and the *level set* costs in the worst case $O(\log(N))$ (insertion in set, each community has at most N neighbors with N the number of vertices) and $O(\log(M)) = O(\log(N))$ running time (the number of distinct edges M is bounded by the square of the number of vertices N^2), respectively.

Merging communities i and j involves an update of the ΔQ matrix and the *level set* for each element of the corresponding rows of the ΔQ matrix. The calculation of each changed value can be achieved in constant time as during the asynchronous parsing it is known whether the other community is linked as well and all other information (community degrees) is stored in a vector. Thus, the total running time contribution of one merging event is $O((d_i + d_j)\log(N))$ with d_k the number of edge starts/ends on vertices of community $k = i,j$. In the worst case all communities are changed in one algorithm round. As the sum over all d_i values is twice the number of distinct edges, the contribution of the merging processes in one algorithm round is at most $O(M \log(N))$. The other steps of one algorithm round are less consumptive: The extraction of pairs belonging to the best l levels can be performed in constant time. The same is true for the update of the degree information. If D is defined as the depth of the dendrogram of communities, at most D algorithm rounds have to be performed. Thus, the running time expectation for

the iterative part is $O(DM\log(N))$ which is identical to the complexity of the classical greedy algorithm [9].

The initialization involves the read-in processes of the edge information (M constant time operations), the degree calculation (part of read-in process), the calculation of the initial modularity (constant time operation on N elements) and finally the generation of the ΔQ matrix and the *level set* at costs $O(M\log(N))$ (M insertions in a set with at most N or M elements, respectively). In the worst case the expected contribution of the initialization to the running time is $O(M\log(N))$.

In the precedent paragraphs we have shown that the MSG greedy algorithm has the total complexity $O(DM\log(N))$. Among the published strategies for modularity optimization the classical greedy algorithm [9] is the fastest [6]. As the MSG shares the worst case expectation for the running time with the classical greedy algorithm, we conclude that the MSG is one of the fastest procedures for modularity optimization.

D. Vertex mover (VM)

To further improve modularity by "adjusting" misplaced vertices, a refinement step called "vertex mover" (VM) is applied upon convergence of the MSG algorithm. In principle, it could also be applied to other modularity optimization procedures. In the VM, the list of vertices is parsed in the order of increasing degree and vertex index (to resolve the degeneracy of multiple vertices with equal degree) and every vertex is reassigned to the neighboring community with maximal modularity improvement. This parsing-and-reassignment procedure is repeated until no modularity improvement is observed.

The VM procedure is similar to the Kernighan-Lin algorithm [10] (applied to modularity optimization in Ref. [11]). In contrast to the Kernighan-Lin algorithm the VM procedure has a perfectly local focus. In other words, instead of repetitively searching for the optimal vertex to reassign, the VM procedure parses the vertices in the aforementioned order and identifies the optimal community for the considered vertex. Furthermore, each reassignment of the VM approach improves modularity. Therefore, the selection of the optimal intermediate partition as in the Kernighan-Lin algorithm is not necessary.

E. VM implementation

The modularity change ΔQ upon reassignment of vertex v from community i to j can be written as

$$\Delta Q = \frac{\text{links}(v \leftrightarrow j) - \text{links}(v \leftrightarrow i)}{L} - \frac{k_v(d_j - d_{i\backslash v})}{2L^2} \qquad (2)$$

with k_v the degree of vertex v, d_j the sum over the degrees of all vertices in community j, $d_{i\backslash v}=d_i-k_v$ the corresponding degree for community i without vertex v, and L the total weight of all edges.

The most time consuming part of the VM is the calculation of the modularity changes upon reassignment of the vertices. Consequently, Eq. (2) reduces this bottleneck to the calculation of weight of the edges connecting the vertex to the neighboring communities. The connectivity information of vertex v is stored in a sparse vector [i.e., a vector of elements (u,w_{vu}) with u a vertex linked to v and w_{vu} the total weight of all edges connecting vertices u and v]. These rows are stored in a vector and form the topology matrix. To determine the total edge weight connecting vertex v with community j the vth row is parsed and for each entry the weight is added to the subtotal edge weight of the corresponding community. To keep access times short a N-dimensional vector (N the number of vertices) is chosen to store the intermediate links$(v \leftrightarrow j)$ results. The optimal reassignment partner for vertex v is the community with smallest index yielding the maximal modularity improvement.

F. Estimation of VM running time

Calculating the modularity changes upon reassignment of one vertex to any neighboring community involves one parse of its edge list supplemented with direct memory access to determine the community affiliation and some constant time operations for the actual modularity calculation. Therefore, the running time contribution of one vertex is proportional to its degree. One algorithm round requires $O(L)=O(\Sigma_i d_i)$ running time. The estimation of the number of needed iterations is not possible as it depends on the quality of the MSG result. In all examples tested by us the running time of the VM was always at least one order of magnitude smaller and less than one minute even for the biggest networks under study.

III. RESULTS

A. Test set of networks

For benchmarking algorithms that optimize modularity the networks commonly used are the collaboration network (coauthorships in cond-mat articles) [12], the graph of metabolic reactions in *caenorhabitis elegans* [13], the email network [14], the network of mutual trust (PGP-key signing) [15,16], the conference graph of college football teams [17], the network of jazz groups with common musicians [18] and the Zachary karate club example [19]. In addition, we include less frequently used examples such as the graph of the metabolic reactions in *Escherichia coli* [20], two different data set describing the protein-protein interactions in S. cerevisiae (budding yeast) [21,22] with labels "PPI" and "yeast." To cover linguistic applications we benchmark the word association network [23] and the graph of the coappearing words in publication titles (co)authored by Martin Karplus [24] who has the third highest h factor [25] among chemists [26]. Further aspects of social webs were incorporated by considering the graph of costarring actors in the IMDB database [27]. Noticeable, the actor network—being the network with the largest number of edges—serves as a proof of concept for such big networks being treatable as well. From computer science we include the internet routing network [28] and the graph of World Wide Web pages [29]. With this selection of networks most currently known application fields of networks are covered. To study the effect of

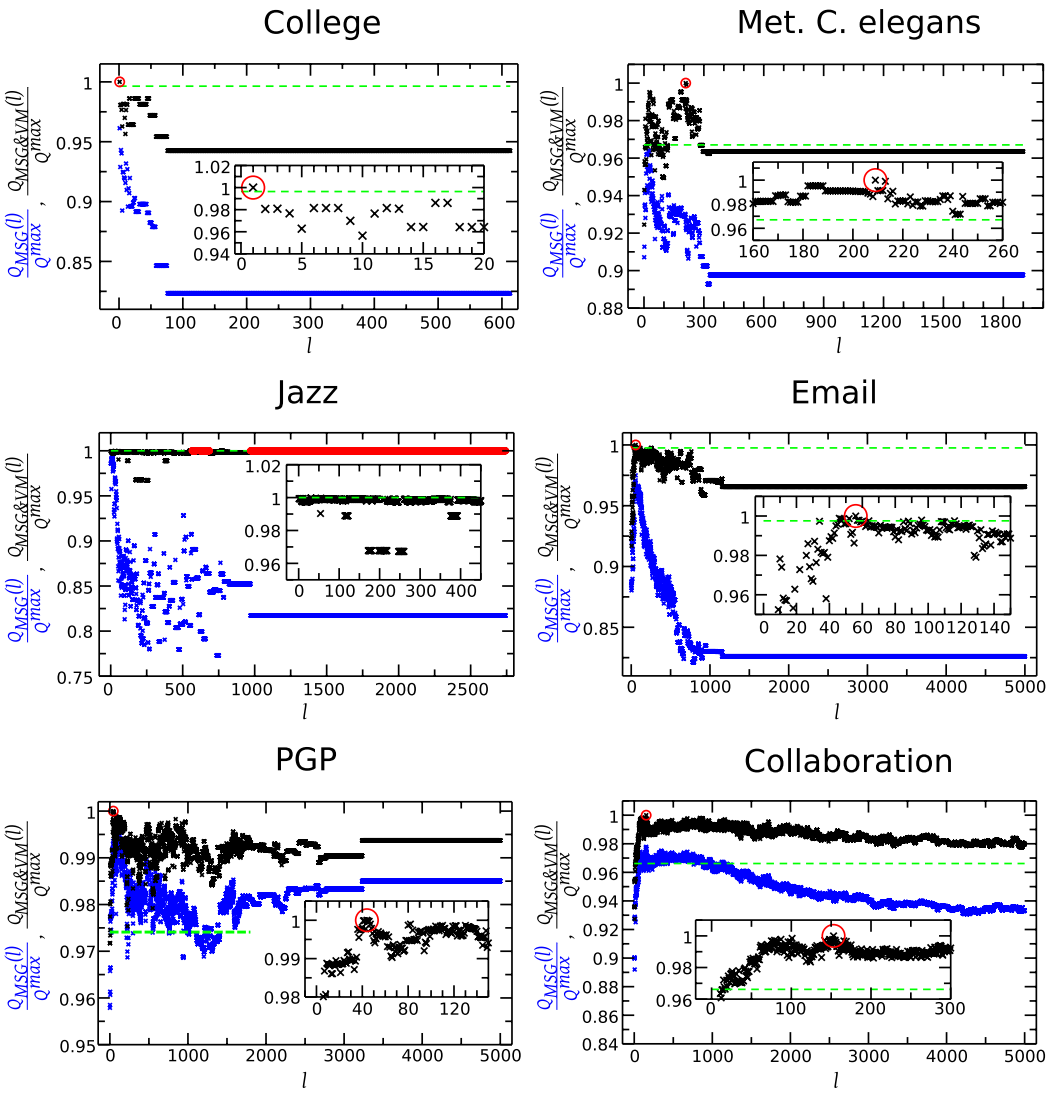

FIG. 2. (Color online) Dependence of MSG modularity value $Q_{\mathrm{MSG}}(l)$ (blue), MSG-VM modularity value $Q_{\mathrm{MSG\text{-}VM}}(l)$ (black) on the level parameter l relative to maximal MSG-VM modularity value $Q^{\max}$. The previously published highest modularity $Q_{\mathrm{pub}}/Q^{\max}$ (dashed green line) is also shown as basis of comparison. The red circles indicate the value of l that yields maximal modularity. A significant number of l values yield higher modularity than the previously published maximal modularity for all but the smallest two networks, i.e., Zachary (not shown) and College. In the latter, only $l=1$ yields a higher modularity than Q_{pub}.

disconnected graphs and weighted networks, we consider in both cases the full network as well as the largest connected component (suffix "CP") and the unweighted variant, respectively. Unless stated otherwise the networks are treated unweighted.

B. Dependence on l and vertex labeling

It is important to investigate the robustness upon the choice of l and to determine the highest modularity values achievable with the MSG-VM algorithm. There is a minor dependence on the value of l (Fig. 2) which changes the MSG-VM modularity by less than 2% for large networks. Moreover, the maximal modularity is obtained with $l<300$ for 14 of the 19 networks (Table I). An empirical formula for the optimal choice of the level parameter will be presented elsewhere.

Noteworthily, for a labeled graph and a chosen level parameter the algorithm is deterministic. To assess the contribution of the labeling, the benchmarking procedure is performed also on hundred copies of the smallest ten networks with permuted vertex labels. This permutation leaves the topology invariant, but modifies the order in which the community pairs are considered. In comparison to the maximal modularity value found for the unscrambled variants a maximal improvement of 0.94% is observed.

TABLE I. Results on real-world examples. Among all tested level parameters (all positive integers smaller than 5000 or the number of edges if smaller) the value l_{opt} yields the highest value of Q for the considered network. N_C is the number of communities found. In most cases, a larger number of communities (larger N_C) is identified by the classical greedy than the MSG-VM extension because the former partitions the network in few large communities and many small communities with less than ten vertices (mostly 2–20 times more small modules identified by greedy than MSG-VM). The MSG-VM approach prevents the condensation into few large modules: The three largest modules contain between 1.5 and 4 times less vertices in the MSG-VM partition than in the greedy partition (not shown). The running time (on a recent laptop) is reported for a single run of the algorithm. The entry "na" indicates that the running time is shorter than 1 s and therefore not displayed. The suffix "CP" points out that only the largest connected component (the "central part") was considered. The acronym "PPI" stands for "protein-protein interaction."

Network				MSG-VM				Greedy		
Name	Ref.	Vertices	Edges	l_{opt}	Q	Time [s]	N_C	Q	Time [s]	N_C
Zachary Karate Club	[19]	34	78	3	0.398	na	4	0.381	na	3
Metabolic *E. coli*	[20]	443	586	6, 8	0.816	na	19	0.811	na	20
College Football	[17]	115	613	1	0.603	na	8	0.556	na	6
Metabolic *C. elegans*	[13]	453	1899	209	0.450	na	8	0.412	na	13
Jazz	[18]	198	2742	566	0.445	na	4	0.439	na	4
Email	[14]	1133	5451	56	0.575	na	10	0.503	na	12
Yeast (PPI, CP)	[21]	2552	7031	35	0.706	na	33	0.675	na	51
M. Karplus weighted	[24]	1166	13423	91	0.316	na	11	0.264	na	18
PPI-CP *S. cerevisiae*	[22]	4626	14801	170	0.545	na	24	0.500	na	38
PPI *S. cerevisiae*	[22]	4713	14846	170	0.546	na	65	0.501	na	81
M. Karplus weighted	[24]	1166	18991	173	0.320	na	13	0.296	na	11
Internet	[28]	11174	23409	278	0.625	8	35	0.584	8	49
PGP-key signing	[15,16]	10680	24340	44	0.878	2	140	0.849	3	195
Word Association (CP)	[23]	7204	31783	71	0.541	4	16	0.452	7	52
Word Association	[23]	7207	31784	97	0.540	3	17	0.465	7	38
Collaboration	[12]	27519	116181	153	0.748	14	82	0.661	103	381
WWW	[29]	325729	1117563	3034	0.939	562	674	0.927	7640	2183
Actor	[27]	82583	3666738	2429	0.543	1722	238	0.470	6288	406
Actor weighted	[27]	82583	4475520	389	0.536	5099	322	0.480	3541	361

C. Performance and running time

The modularity values obtained with the MSG-VM approach are listed in Table II. For five of the seven networks considered here the MSG-VM algorithm finds solutions with modularity higher than previously published. Only for the Zachary Karate network the MSG-VM procedure yields a smaller modularity value. For the jazz network a solution with the identical Q value is obtained. For the networks without published modularity values we compare the optimal values obtained by the MSG-VM algorithm with the classical greedy algorithm for modularity optimization as introduced by Newman [5] in Table I. We observe that the MSG-VM algorithm outperforms the original greedy algorithm significantly.

The running time estimations in Secs. II C and II F are based on a worst case scenario. To investigate the running time behavior on real-world examples, we compare the running times of the classical greedy variant and the MSG-VM algorithm in Table I. These data show that given the appropriate level parameter choice the MSG-VM algorithm is in almost all cases faster than the classical greedy algorithm and, at the same time, reaches a higher value of modularity.

IV. CONCLUSIONS

To prevent premature condensation into few large communities the greedy algorithm for modularity optimization has been extended by a procedure for simultaneous merging of more than one pair of communities at each step. Further-

TABLE II. Comparison of maximal value of modularity obtained by the MSG-VM algorithm $Q_{max}^{MSG\text{-}VM}$ with previously published results Q_{pub}. The highest published value was extracted from the referenced paper ("Source") where it has been calculated by the "Method" whose reference is listed in the last column.

Network	$Q_{max}^{MSG\text{-}VM}$	Q_{pub}	Source	Method
Zachary Karate Club	0.398	0.419	[11]	[11]
College Football	0.603	0.601	[17]	[17]
Metabolic *C. elegans*	0.450	0.435	[11]	[11]
Jazz	0.445	0.445	[11]	[3]
Email	0.575	0.574	[11]	[3]
PGP-key signing	0.878	0.855	[11]	[11]
Collaboration	0.748	0.723	[11]	[11]

more, this "multistep" greedy variant has been combined with a simple vertex-by-vertex *a posteriori* refinement. On seven networks with previously published modularity values the MSG-VM algorithm combination outperforms all other frequently used, generic techniques except for the smallest of the seven examples. In addition, a single run of the MSG-VM algorithm requires similar computer time as the greedy algorithm. In most cases less than 10 independent (i.e., embarrassingly parallel) runs of MSG-VM are required to obtain a modularity within 1% of the highest value because an empirical formula has been derived for the appropriate choice of the optimal step width [30]. Therefore, the MSG-VM algorithm is an efficient tool to find network partitions with high modularity [31].

ACKNOWLEDGMENTS

The authors thank Stefanie Muff and Francesco Rao for helpful discussions. Christian Bolliger, Thorsten Steenbock, and Dr. Alexander Godknecht are acknowledged for maintaining the Matterhorn cluster where most of the parameter studies were performed. We are thankful to Drs. Arenas, Barabási, Gleiser, and Newman for providing the network data. This work was supported by a Swiss National Science Foundation grant to A.C.

[1] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[2] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, e-print arXiv:physics/0608255.
[3] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[4] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[5] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[6] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech. **2005**, P09008 (2005).
[7] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[8] J. M. Kumpula, J. Saramäki, K. Kaski, and J. Kertész, Eur. Phys. J. B **56**, 41 (2007).
[9] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[10] B. Kernighan and S. Lin, Bell Syst. Tech. J. **49**, 291 (1972).
[11] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **103**, 8577 (2006).
[12] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[13] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, Nature (London) **407**, 651 (2000).
[14] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103(R) (2003).
[15] X. Guardiola, R. Guimerà, A. Arenas, A. Díaz-Guilera, D. Streib, and L. A. N. Amaral, e-print arXiv:cond-mat/0206240.
[16] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).
[17] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[18] P. Gleiser and L. Danon, Adv. Complex Syst. **6**, 565 (2003).
[19] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1974).
[20] H. Ma and A.-P. Zeng, Bioinformatics **19**, 270 (2003).
[21] N. J. Krogan *et al.*, Nature (London) **440**, 637 (2006).
[22] V. Colizza, A. Flammini, A. Maritan, and A. Vespignani, Physica A **352**, 1 (2005).
[23] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, Behav. Res. Methods Instrum. Comput. **36**, 402 (2004).
[24] P. Schuetz and A. Caflisch, the network of words in the titles of Martin Karplus' publications (unpublished).
[25] J. E. Hirsch, Proc. Natl. Acad. Sci. U.S.A. **102**, 16569 (2005).
[26] P. Ball, Nature (London) **448**, 737 (2007).
[27] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[28] Internet Network. Undirected, unweighted network of the Internet at the Autonomous System level from data collected by the Oregon Route Views Project (http://www.routeviews.org/) in May 2001, where vertices represent Internet service providers and edges connections among them. The file reports the list of connected pairs of nodes.
[29] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[30] Ph. Schuetz and A. Caflisch (in preparation).
[31] The code is available at http://www.biochem-caflisch.uzh.ch/communitydetection/

Chapter 4

Multistep greedy algorithm identifies community structure in real-world and computer-generated networks.

P. Schuetz and A. Caflisch

[*Phys. Rev. E*, **2008**, *78*, 026112]

Multistep greedy algorithm identifies community structure in real-world and computer-generated networks

Philipp Schuetz* and Amedeo Caflisch†
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
(Received 18 April 2008; published 20 August 2008)

We have recently introduced a multistep extension of the greedy algorithm for modularity optimization. The extension is based on the idea that merging l pairs of communities ($l>1$) at each iteration prevents premature condensation into few large communities. Here, an empirical formula is presented for the choice of the step width l that generates partitions with (close to) optimal modularity for 17 real-world and 1100 computer-generated networks. Furthermore, an in-depth analysis of the communities of two real-world networks (the metabolic network of the bacterium *E. coli* and the graph of coappearing words in the titles of papers coauthored by Martin Karplus) provides evidence that the partition obtained by the multistep greedy algorithm is superior to the one generated by the original greedy algorithm not only with respect to modularity, but also according to objective criteria. In other words, the multistep extension of the greedy algorithm reduces the danger of getting trapped in local optima of modularity and generates more reasonable partitions.

DOI: 10.1103/PhysRevE.78.026112 PACS number(s): 89.75.Fb, 05.10.−a, 89.75.Kd, 89.75.Hc

I. INTRODUCTION

The coarse-grained organization of many real-world networks manifests itself in a natural divisibility of the vertices into modules (or communities). A community is a set of vertices that are more connected among each other than with vertices of other communities. Community structure has been reported for social networks [1,2], metabolic networks [3–5], and protein-folding networks [6–10]. Several procedures have been developed to partition a network into modules. Often applied are techniques that rely on the optimization of a scoring function called *modularity* [11]. This assessment function compares the fraction of edges within a module with its expectation value in the case of randomly connected vertices with equal degree sequence. The modularity is defined as

$$Q = \sum_{i=1}^{N_C} \left[\frac{I(i)}{L} - \left(\frac{d_i}{2L} \right)^2 \right], \tag{1}$$

with $I(i)$ being the weights of all edges linking vertices of community i, d_i the sum over all vertex degrees in module i, L the total edge weight, and N_C the number of communities. The optimization of modularity has been proven to be a *NP*-hard problem [12]. Thus, heuristic techniques such as extremal optimization [13], simulated annealing [4], and the greedy algorithm [14] have been developed and applied to find partitions with high modularity. Because of the global character of modularity [i.e., in Eq. (1) the connectivity and degree of each community are compared with the edge weight of the whole network], it has been shown that modules smaller than a certain scale cannot be resolved [15]. In other words, small communities are amalgamated with others instead of being detected autonomously. A higher-resolution variant of modularity, called *localized* modularity, addresses the limit on the detectable community size [5].

Recently, we have introduced a multistep extension of the greedy algorithm (MSG) and combined it with a simple vertex-by-vertex refinement procedure [vertex mover (VM)] [16]. The essential idea of the MSG algorithm is to promote the simultaneous merging of several pairs of communities to prevent premature trapping in a local optimum of modularity. Given an appropriate choice of the step width l, the MSG-VM algorithm finds partitions with high modularity in short running time. Our implementation of the MSG-VM algorithm [16,17] has the same scaling behavior as the efficient version of the greedy algorithm [18], which has the smallest complexity among the commonly used community-detection algorithms [19]. Note that the running time of both the MSG-VM algorithm [16] and the greedy algorithm [18] is $O(DL \log N)$ with L, N, and D the number of edges, vertices, and the depth of the dendrogram describing the community structure, respectively. For a sparse network with $L \sim N$ and $D \sim \log N$, the scaling is essentially linear $O(N \log^2 N)$.

In this paper, we derive an empirical formula for predicting optimal l values—i.e., values of the step width that yield a modularity very close to the highest value achievable by the MSG-VM algorithm. Furthermore, for two real-world networks having each an inherent partition into substructures, we compare the community structures identified by the original greedy and the MSG-VM algorithm. These real-world examples are the metabolic network of *E. coli* and the graph of coappearing words in the titles of publications coauthored by Martin Karplus, the most cited theoretical chemist. The inherent substructures of the former are the metabolic pathways, while the inherent substructure of the network of Karplus' paper titles are the sets of words predominantly used in research subfields in theoretical and computational chemistry. These two examples illustrate that the

*schutz@bioc.uzh.ch
†FAX: +41 44 635 68 62 caflisch@bioc.uzh.ch

MSG-VM algorithm detects the underlying substructures more accurately than the original greedy algorithm.

II. METHODS

A. Multistep greedy and vertex mover algorithms

The MSG algorithm optimizes modularity by an iterative procedure in which multiple pairs of communities are merged at each iteration. This multistep approach is a significant extension with respect to the original greedy algorithm [14], in which only the pair of communities that improves modularity most is merged in each iteration. A pseudocode description of the MSG algorithm is

Initialization:
Each vertex is a community;
Calculate matrix ΔQ whose elements are the modularity changes upon merging of module pair (i,j);
Iteration:
while pair (i,j) with $\Delta Q_{ij}>0$ exists **do**
  for all triplets $(i,j,\Delta Q_{ij})$ of ΔQ, parsed w.r.t. decreasing ΔQ_{ij} and increasing (i,j)
  do
    if $\left\{ \begin{array}{c} \Delta Q_{ij} > 0 \text{ in best } l \text{ values in } \Delta Q \text{ - matrix} \\ i \text{ and } j \text{ unchanged in iteration} \end{array} \right\}$
    then
      MergeCommunities(i,j)
    end if
  end for
end while

Algorithm 1: Flowchart of the MSG procedure. Details of the efficient merge of two communities and the calculation of the modularity change matrix are presented in [16].

Note that the step width l influences the number of merged pairs (but is not necessarily identical to it); furthermore, l is kept constant during an MSG run (for more details, the reader is referred to the original publication [16]).

Applied upon convergence of the MSG algorithm the VM procedure improves modularity by "adjusting" misplaced vertices. The VM procedure parses the vertex list in ascending vertex degree and index order and checks for each vertex whether a reassignment to one of the neighboring communities yields a modularity improvement [16].

B. Networks

All networks in this article are treated undirected and unweighted.

1. Real-world networks

The real-world networks are the same as in [16] and are listed in Table I. Sociological applications are included with the Zachary karate club example [20], the conference graph of college football teams [21], the graph of jazz groups with common musicians [2], the network of mutual trust (PGP-key signing) [27,28], the collaboration network (coauthorships in cond-mat articles) [1], and the graph of costarring actors in the IMDB database [31]. Network applications in biochemistry are covered by the graph of metabolic reactions in the nematode *Caenorhabditis elegans* [22] and the bacterium *Escherichia coli* [3] as well as two different data sets describing the protein-protein interactions in *Saccharomyces cerevisiae* (budding yeast) [24,25] with labels "PPI" and "yeast." Linguistic applications are covered by the Word Association network [29] and the graph of the coappearing words in titles of publications (co)authored by Martin Karplus [16,17] who has the third highest h factor [33] among chemists [34]. From computer science the internet routing network [26] and the graph of WWW pages [30] are included. The effects of disconnected graphs are considered by including the full network as well as its largest connected component (LCC).

2. Computer-generated networks

A total of 1100 computer-generated networks were used for an in-depth assessment of the empirical formula for the prediction of optimal values of l (Table II). The networks in $GN_{1,2,3}$ consist of 128 vertices organized in four equally sized communities [21]. The cohesion of the vertices within a module is controlled by a parameter called z_{out} which determines the average number of edges connecting vertices of different modules. To consider clearly formed and loosely coupled modules the z_{out} parameter is chosen uniformly from the second smallest to the highest value. Among the sets $GN_{1,2,3}$, the number of edges is varied to assess the effect of different values of average degree.

The remaining test cases are larger networks, which have no imposed community structure and a heterogeneous distribution of the vertex degrees and community sizes (confer Table 1 in the supplementary material [32]). A recent study, published after the submission of this work, has emphasized the importance of this heterogeneity for testing community-detection algorithms on severe benchmarks [35]. To foster a "spontaneous" formation of modules a vertex degree distribution is imposed. The network is generated by choosing a number of vertices at random (uniform distribution), assigning edge end points to each vertex according to the degree distribution and joining the edge endpoints at random. To examine the effect of different degree distributions, an exponential distribution is used for the networks in SED (small networks with exponential degree distribution) and a linear distribution is imposed on the networks in SLD and LLD. All networks in LLD have at least 300 vertices. After generation, the networks in SED, SLD, and LLD are projected onto the largest connected component and treated as unweighted.

III. RESULTS

It is helpful to recall here that L is the number of edges and l_{opt} is the value of the step width that yields the highest MSG-VM modularity (among all tested values of step width). The MSG-VM algorithm is applied on each real-world network using every integer $l<\min\{5000,L\}$. The modularity values before and after the VM application are

TABLE I. Properties of real-world networks and comparison of MSG-VM runs using l as in Eq. (2) or picked at random. The column "Q_{opt}" lists the maximal value of modularity obtained by running MSG-VM for all values of l smaller than $\min\{5000, L\}$ (where L is the number of edges). The column "Q_{pred}" lists the MSG-VM modularity obtained using Eq. (2) to determine the step width. The columns "$\langle Q_{rand}\rangle$" and "$\langle Q_{rand}^{l<1.5\sqrt{L}}\rangle$" show the expectation value for the MSG-VM modularity when six values of l are picked randomly from a uniform distribution in the range $1 \leqslant l \leqslant \min\{5000, L\}$ and $1 \leqslant l \leqslant 1.5\sqrt{L}$, respectively. The expectation value is estimated by averaging, over 1000 samples, the highest modularity obtained using six values of l (details are given in Sec. VII of the supplementary material [32]). Six values of l are picked randomly for each sample because six values were used to determine Q_{pred}: the four values of l calculated by Eq. (2) and the two integers adjacent to the best of these four. Values of $\langle Q_{rand}\rangle$ and $\langle Q_{rand}^{l<1.5\sqrt{L}}\rangle$ higher than the corresponding Q_{pred} are in italics. The acronym LCC stands for "largest connected component."

				MSG-VM with Optimal l		MSG-VM with l from Eq. (2)	MSG-VM with Random l	
Network	Ref.	Vertices	Edges (L)	$l_{opt}/\sqrt{L}$	Q_{opt}	Q_{pred}	$\langle Q_{rand}\rangle$	$\langle Q_{rand}^{l<1.5\sqrt{L}}\rangle$
Zachary Karate Club	[20]	34	78	0.34	0.398	0.398	0.391	0.398
Metabolic *E. coli*	[3]	443	586	0.25	0.816	0.816	0.813	0.816
College Football	[21]	115	613	0.04	0.603	0.595	0.579	*0.596*
Metabolic *C. elegans*	[22]	453	1899	4.80	0.450	0.447	0.439	0.445
Jazz	[2]	198	2742	10.81	0.4451	0.4447	*0.4451*	*0.4448*
Email	[23]	1133	5451	0.76	0.575	0.575	0.564	0.574
Yeast (PPI, LCC)	[24]	2552	7031	0.42	0.706	0.705	0.693	0.702
M. Karplus	[16,17]	1167	13423	0.79	0.316	0.311	0.306	0.311
PPI *S. cerevisiae* (LCC)	[25]	4626	14801	1.40	0.545	0.544	0.531	0.543
PPI *S. cerevisiae*	[25]	4713	14846	1.40	0.546	0.546	0.532	0.545
Internet	[26]	11174	23409	1.82	0.625	0.619	0.615	0.618
PGP-key signing	[27,28]	10680	24340	0.28	0.878	0.876	0.873	0.876
Word Association (LCC)	[29]	7204	31783	0.40	0.541	0.536	0.528	0.536
Word Association	[29]	7207	31784	0.54	0.540	0.537	0.527	0.536
Collaboration	[1]	27519	116181	0.45	0.748	0.746	0.743	0.744
WWW	[30]	325729	1117563	2.87	0.939	0.936	*0.937*	*0.937*
Actor	[31]	82583	3666738	1.27	0.543	0.536	*0.537*	*0.539*

recorded. For the computer-generated networks all integer values $l < 10\sqrt{L}$ have been tested (the $\sqrt{L}$ scaling is rationalized in the next subsection).

A. Dependence of l on network properties

The correlation between the optimal step width l_{opt} and several topological properties was calculated. The following properties or powers thereof were used: number of vertices and edges, highest degree, average degree, standard deviation of degree, average of power 1, 2, or 3 of the clustering coefficient, and average and standard deviation of the differences in clustering coefficient values or degree of linked vertices. The highest correlation was observed for $\sqrt{L}$ (0.7728, correlation coefficients of other properties are listed in the supplementary material [32]).

This empirical result is consistent with the $\sqrt{L}$ dependence of the number of communities yielding maximal modularity as recently demonstrated for one class of networks [15]. In fact, a close inspection of the MSG algorithm shows that the step width l determines the number of communities formed during the first 1%–5% of the iterations (the number of iterations is strongly dependent on the network topology). Each module in the final solution has to be nucleated as early as possible and therefore l to be chosen according to the expected number of communities.

1. Optimal prefactor for computer-generated networks

To determine the prefactor α in the $\sqrt{L}$-scaling law the computer-generated networks introduced in Sec. II B 2 are examined first. This choice is due to their multitude (1100 networks) and their lack of overlapping condensed structures [i.e., few (almost) complete subgraphs sharing vertices] as consequence of the construction principle. First, we observe that for 97 of the 1100 networks the MSG-VM modularity does not depend on l. Further, for each value of α the MSG-VM modularity is averaged over all networks of the same type $\bar{Q}_{\mathrm{MSG-VM}}(\alpha) = \frac{1}{N_S}\Sigma_{i\in S}\frac{Q^i_{\mathrm{MSG-VM}}(\lfloor \alpha\sqrt{L_i}\rfloor)}{\max_l(Q^i_{\mathrm{MSG-VM}}(l))}$, where S is the type of networks, N_S is the number of networks of type S, $\lfloor . \rfloor$ is the floor function, and L_i is the number of edges in network i. All α profiles peak for $0.2 < \alpha < 0.3$ and show a similar behavior (Fig. 1). The α profiles averaged over all computer-generated networks peak at $\alpha = 0.251$. [It is legitimate to consider the average because for each α the histogram of $\frac{Q^i_{\mathrm{MSG-VM}}(\lfloor \alpha\sqrt{L_i}\rfloor)}{\max_l(Q^i_{\mathrm{MSG-VM}}(l))}$ (i indexing the network realizations) follows an unimodal distribution with an additional peak at

TABLE II. Properties of computer-generated networks. The networks in the GN_i (Girvan-Newman) sets ($i=1,2,3$) possess an imposed four community structure where z_{out} controls the average number of edges connecting two different modules [21]. For the networks of type SED (small networks with exponential degree distribution), SLD (small networks with linear degree distribution), and LLD (large networks with linear degree distribution) a degree distribution has been prescribed to foster the formation of communities.

Type	No. of realizations	Vertices	Edges	Remarks
GN_1	100	128	1024	$z_{out}=3-16$
GN_2	100	128	512	$z_{out}=2-8$
GN_3	100	128	2048	$z_{out}=2-32$
SED	300	11–976	10–19247	Exp. deg. distr.
SLD	200	19–3777	43–78741	Linear deg. distr.
LLD	300	309–4278	1523–342940	Linear deg. distr.

1.0 originating from the degeneracy of l^i_{opt}.] Excluding the additional peak, the highest normalized modularities are still observed for $0.2<\alpha<0.3$. Remarkably, the degeneracy of l^i_{opt} [i.e., the number of networks with $Q^i_{MSG-VM}(\lfloor\alpha\sqrt{L_i}\rfloor) = \max_l(Q^i_{MSG-VM}(l))$] is highest for $0.18<\alpha<0.26$. A leave-N-out procedure (confer supplementary material [32] for details) provides evidence that $\alpha=0.251$ would have been (close to) optimal also for another selection of networks. The application of the MSG-VM algorithm with step width $\lfloor 0.251\sqrt{L}\rfloor$ yields 97.6% of the highest MSG-VM modularity averaging over all computer-generated networks (98% if median is calculated).

2. Comparison of empirical formula with random selection of step width

If a step width value is selected at random among $l<\min\{L,5000\}$ (all tested values), the MSG-VM algorithm is expected to yield 93.4% of the highest MSG-VM modularity on average over all computer-generated networks [the expectation value is equal to the arithmetic mean over all $Q_{MSG-VM}(l)$ values]. An in-depth analysis (details given in the supplementary material [32]) shows that $l_{opt}<1.5\sqrt{L}$ for 92.6% of all computer-generated networks. If a step width value smaller than $1.5\sqrt{L}$ is chosen at random, the expectation value of the MSG-VM modularity raises to 95.9% of its highest value (average over all computer-generated networks). Thus, the empirical formula $l=0.251\sqrt{L}$ performs 4.3% better (of a maximum of 6.6%) than a value of step width picked at random if all tested values are considered. If the reduced test set $l<1.5\sqrt{L}$ is used, the empirical formula performs 1.7% better than a value of step width picked at random (4.1% maximal improvement). More precisely, for 85.5% of the networks the MSG-VM modularity with $l=0.251\sqrt{L}$ is higher than the one with l picked at random and the average improvement for these networks is 2.4%.

To account for limited sampling the prefactor $\alpha=0.25$ is assumed to be optimal for the computer-generated networks (the prefactors 0.251 and 0.25 can be considered identical as the real to integer conversion yields the same value of l for networks with $L<10^6$).

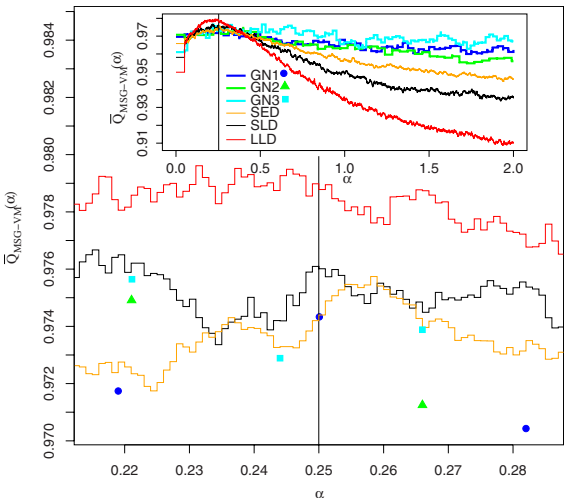

FIG. 1. (Color online) Dependence of Q_{MSG-VM} on the $\sqrt{L}$ prefactor α for the computer-generated networks. The averages are taken separately for each type of computer-generated networks. The vertical black line denotes $\alpha=0.25$, which is the value suggested in Eq. (2). The parameter range for α has been discretized to multiples of 0.001 to simplify the calculations.

3. Application to real-world networks

In comparison to computer-generated graphs, real-world networks are endowed with more condensed substructures. Therefore, a different scaling behavior than for the computer-generated networks is possible. To improve statistics and reduce spurious effects due to vertex-labeling artifacts (a value of step width yields a high MSG-VM modularity as it profits exclusively from the "right" parsing of the vertices), 100 copies of the smallest 10 real-world networks are created with permuted vertex labelings (details are presented in the supplementary material [32]). For each copy the influence of l is tested as described in Sec. III. Except for the College Football and Email networks all $\bar{Q}_{MSG-VM}$ profiles (confer Sec. III A 1 for the definition) averaged over the scrambled variants are observed to peak for values of step width equal or very close to

$$l=\lfloor\alpha\sqrt{L}\rfloor \quad (\alpha=0.25,0.5,0.75,1) \tag{2}$$

(supplementary material [32]). The MSG-VM modularity deviates at most by 1.47% from the maximal value (Table I). Moreover, for 13 of the 17 networks the deviation is smaller than 0.94%. In comparison to the effect of permuted vertex labels this deviation is of the same order of magnitude and thus negligible (details given in the supplementary material [32]).

To further assess the predictive power of Eq. (2), the MSG-VM modularity obtained with l as in Eq. (2) is compared with a random selection of the step widths. Because of the real to integer conversion induced by the floor function, an integer adjacent to $\lfloor\alpha\sqrt{L}\rfloor$ might be optimal. Therefore, not only the four values of step width as in Eq. (2) are tested, but

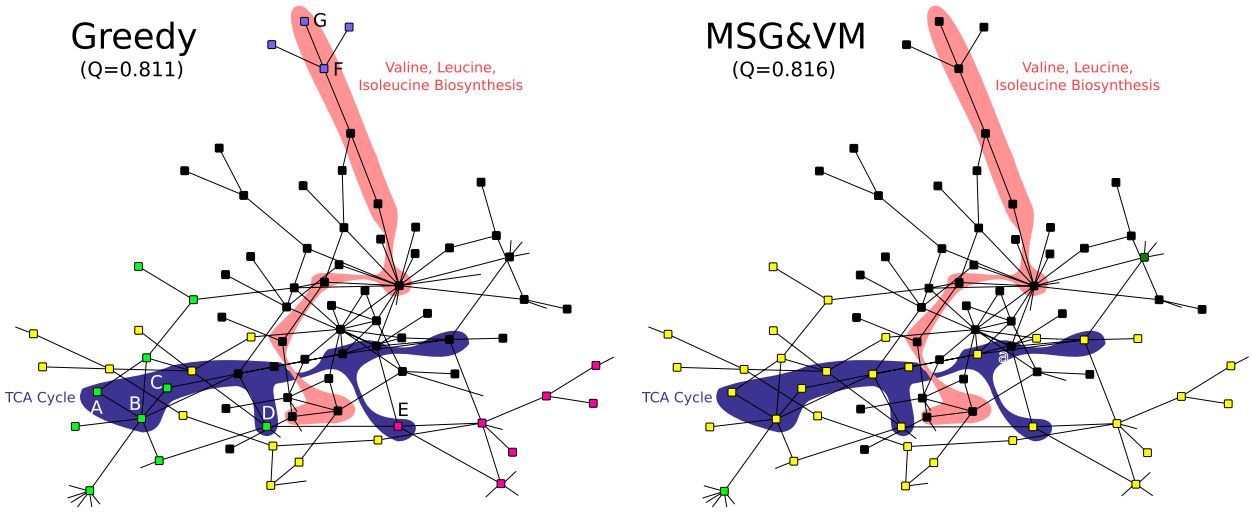

FIG. 2. (Color online) Clusterization of the metabolic network of *E. coli* and accuracy of pathway identification. Two exemplary pathways as taken from the KEGG database [36,37] (pathways MAP00020 for "TCA cycle" and MAP00290 for "Valine, Leucine, Isoleucine Biosynthesis") are highlighted by the colored areas. An excerpt of the network is shown here while the full network is in the supplementary material [32]. The misassigned vertices are indicated by letters; they are a=(S)-Malate for MSG-VM, and for the original greedy: A=3-Carboxy-1-hydroxypropyl-ThPP, B=2-Oxoglutarate, C=Oxalosuccinate, D=Succinate, E=Fumarate, F=2-Oxoisovalerate, and G =Valine.

also the two integers adjacent to the best of them. For a fair comparison the same number of trials is allowed in the random experiment. For 14 out of 17 networks the MSG-VM modularity value with l as in Eq. (2) is higher or equal than for the corresponding random experiment (Table I). Therefore, one can conclude that the empirical formula (2) yields a step width that results in (close to) optimal modularity, and therefore only six runs of the MSG-VM algorithm are required.

B. Quality of MSG-VM network partition

Previously, the performance of the MSG-VM algorithm in optimizing modularity has been shown on 19 real-world networks [16]. Here, an in-depth analysis of two examples provides evidence that the MSG-VM algorithm gathers vertices in groups that represent substructures (identified by other means) more accurately than the greedy algorithm.

1. Metabolic network of E. coli

The network of metabolic reactions in the bacterium *E. coli* is extracted from the KEGG database (data set "Escherichia coli K-12 MG1655") with additional refinement by Ma and Zeng [3] and projected on the largest connected component. Furthermore, chains of vertices with degree 1 or 2 are reduced to one single vertex (to reduce spurious effects of unnaturally splitted chains). Each vertex is assigned to between zero and eight out of 11 metabolic pathways with an average of 1.51 ± 0.99.

Eleven communities are identical in the MSG-VM partition (which consists of 19 communities and has $Q=0.816$) and the partition obtained with the greedy algorithm (20 communities, $Q=0.811$). To assess the quality of pathway detection, we employ the measure $P=\Sigma_i \frac{P_i}{N_i}$ (adopted from [5]), with P_i the number of vertex pairs in community i that share at least one pathway and N_i the number of pairs of vertices with equal community affiliation. The MSG-VM partition has $P_{\mathrm{MSG-VM}}=0.60$, which is better than the partition obtained with the original greedy algorithm ($P_{\mathrm{greedy}}=0.58$). The improved pathway identification is illustrated by an excerpt of the network in Fig. 2 (vertices in the 11 modules which are identical in both partitions are removed for visibility reasons). Two central pathways (classification according to KEGG database) are highlighted by colored areas. In the MSG-VM solution the vertices of each pathway belong to separate modules except for "(S)-Malate." This metabolite has more connections to vertices assigned to the "Amino Acid Metabolism" than to those of the "Carbohydrate Metabolism" (the "TCA cycle" is associated to the latter) and thus, a separation is meaningful. On the other hand, the metabolites misclassified by the original greedy algorithm are "2-Oxo-glutarate" (B), "3-Carboxy-hydroxypropyl-ThPP"(A), and "Oxalosuccinate"(C). The last two belong only to the "TCA cycle" pathway, whereas "2-Oxo-glutarate" is part of several pathways and therefore can also be attributed to other communities. Furthermore, the separation of the blue vertices in the "Valine, Leucine, Isoleucine Biosynthesis" pathway is peculiar as the overlapping pathway "pantothenate and CoA biosynthesis" is contracted to one vertex (the vertex right to "F" and "G"). The metabolites "F" and "G" are the educts in the "pantothenate and CoA biosynthesis" pathway. If a unique assignment has to be made, an attribution to the "Valine, Leucine, Isoleucine Biosynthesis" pathway is more reasonable. The last differences of the greedy partition to the MSG-VM solution are "Succinate" (D) and "Fumarate" (E) which are as "(S)-Malate" (a) part of multiple different metabolic processes and therefore may be attributed to multiple pathways. To summarize, of eight assignments differing between MSG-VM and original greedy

TABLE III. The five largest communities as identified by the MSG-VM algorithm in the network of words in the titles of M. Karplus' papers. These five communities account for 81% of the vertices in the network. Unspecific words (e.g., "study" and "theory" with degree 291 and 234, respectively) were taken into account for the clusterization, but are not listed in this table.

		Most frequent words		Number of titles with any of the words in community	Description
Rank	Vertices	Degree	Word		
1	220	407	Protein	442	Molecular dynamics (of proteins)
		318	Simulation		
		269	Molecular-dynamics		
2	184	290	Structure	330	Three-dimensional structures
		123	Peptide		
		97	Inhibitor		
3	162	269	Model	335	Molecular modelling, molecular mechanics
		178	Energy		
		169	Function		
4	162	159	Molecule	306	Quantum mechanics, free-energy calculation
		154	Free-energy		
		144	Potential		
5	116	212	Reaction	205	Chemical reaction, kinetics, and solvation
		154	Solution		
		101	Solvation		

algorithm (in the excerpt of the network shown in Fig. 2), none was misplaced by the MSG-VM algorithm, whereas the greedy algorithm misplaced two metabolites (two further examples of incomplete detection of pathways by the original greedy algorithm are shown in the supplementary material [32]).

2. Network of words in titles of Karplus' publications

Martin Karplus is one of the most productive and most cited chemists (78091 citations as of 3 July 2008). As second example we analyze the community structure of the graph of words coappearing in the titles of the 719 publications coauthored by Karplus between 1947 and 2004 [16,17]. The words with highest degree in the five largest (according to number of words) communities are shown in Table III.

The following two examples provide evidence for the superiority of the MSG-VM partition (11 communities, $Q=0.316$) with respect to the partition obtained by the original greedy algorithm (18 communities, $Q=0.264$). The words "reaction" (degree 212), "hydrolysis" (73), "rate" (69), "enzyme" (57), "catalysis" (54), and "kinetics" (54) are appropriately grouped in a single community by the former, while they are spread in the four largest (according to the number of words) communities by the latter. Another example of superiority of the MSG-VM partition is the community with the words "molecule" (159), "atom" (91), and "bond" (87), which are spread over the three largest communities by the greedy algorithm. These two examples show that the main advantage of the MSG-VM algorithm is that the simultaneous emergence of several communities hinders the spurious coalescence into few large communities observed for the original greedy algorithm.

IV. CONCLUSIONS

The performance of the MSG procedure, a multistep extension of the greedy algorithm, was analyzed on 1100 computer-generated networks of heterogeneous sizes and degree distributions and 17 real-world networks. Several powers of topological properties (e.g., average degree, clustering coefficient, etc.) were tested as prediction formulas for the optimal step width l. The empirical formula $l=\lfloor\alpha\sqrt{L}\rfloor$ (L total edge weight; $\alpha=0.25,0.5,0.75,1$) outperforms all others and yields a higher modularity value than a random picking of the step width for 85.5% of the computer-generated networks and 14 of 17 real-world examples. For these 14 real-world networks, the modularity optimized by MSG-VM algorithm using only six values of l ($l_1=\lfloor 0.25\sqrt{L}\rfloor$, $l_2=\lfloor 0.5\sqrt{L}\rfloor$, $l_3=\lfloor 0.75\sqrt{L}\rfloor$, $l_4=\lfloor 1.0\sqrt{L}\rfloor$, and $l_{5,6}=l_{\max}\pm 1$ with $l_{\max}$ the step width among $l_{1,\ldots,4}$ that yields the highest modularity) is larger than 99% of the highest value achievable by exhaustive testing of all step widths (i.e., $1\leqslant l\leqslant L$). This deviation is on the order of the fluctuations observed when the parsing order of the vertices is changed. In addition, for 92.6% of the computer-generated and 13 of 17 real-world networks the optimal value of the step width is smaller than $1.5\sqrt{L}$.

To assess the quality of the community identification two real-world examples (the network of metabolic reactions in *E. coli* and the graph of coappearing words in titles of publications coauthored by M. Karplus) were examined in-depth and the modular structure obtained from the application of the MSG-VM and greedy algorithms was compared. For the metabolic network the original greedy algorithm splits two exemplary pathways ("TCA cycle" and "Valine, Leucine, Isoleucine Biosynthesis") in multiple parts with seven misplaced vertices. Two of these vertices are not part of another pathway and therefore are wrongly assigned by the original greedy algorithm. For the MSG-VM solution only one metabolite is misplaced which can be attributed to the three pathways in which this metabolite is involved. Furthermore, an objective criterion (the conditional probability that two vertices in the same module share at least one pathway) sup-

ports these exemplary observtions. For the "M. Karplus" network the partition obtained by the original greedy algorithm has three very large modules in which words of distinct research fields are inappropriately mixed. Moreover, subsets of words belonging to the same topic are erroneously split (e.g., "atom," "molecule," and "bond" are split in the three largest modules). On the other hand, the MSG-VM procedure more accurately groups subsets of words belonging to individual research topics.

In conclusion, the MSG-VM algorithm is one of the fastest and most accurate procedures for modularity optimization currently available because it scales as $O(N \log^2 N)$ for a sparse network (N the number of vertices) [16]. Therefore, a single run is faster than previously published approaches [19], and only six independent runs are required using Eq. (2) to determine the step width [17].

ACKNOWLEDGMENTS

We thank Stefanie Muff for helpful comments on the manuscript. Christian Bolliger, Thorsten Steenbock, and Dr. Alexander Godknecht are acknowledged for maintaining the Matterhorn cluster where most of the presented parameter studies were performed. For providing the data sets we are thankful to Dr. A. Arenas, Dr. A. L. Barabasi, Dr. P. Gleiser, Dr. H. Ma, and Dr. M. E. J. Newman [17]. This work was supported by a Swiss National Science Foundation grant to A.C.

[1] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).
[2] P. Gleiser and L. Danon, Adv. Complex Syst. **6**, 565 (2003).
[3] H. Ma and A.-P. Zeng, Bioinformatics **19**, 270 (2003).
[4] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895 (2005).
[5] S. Muff, F. Rao, and A. Caflisch, Phys. Rev. E **72**, 056107 (2005).
[6] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[7] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **102**, 18914 (2005).
[8] D. Gfeller, P. D. L. Rios, A. Caflisch, and F. Rao, Proc. Natl. Acad. Sci. U.S.A. **104**, 1817 (2007).
[9] S. Muff and A. Caflisch, Proteins **70**, 1185 (2008).
[10] S. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).
[11] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[12] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, e-print arXiv:physics/0608255.
[13] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[14] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[15] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[16] P. Schuetz and A. Caflisch, Phys. Rev. E **77**, 046112 (2008).
[17] The source code of the MSG-VM algorithms, the awk scripts to generate the SED, SLD, and LLD networks, as well as the network of words in the titles of Martin Karplus' publications are available at http://www.biochem-caflisch.uzh.ch/communitydetection
[18] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[19] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech.: Theory Exp. 2005, P09008
[20] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1974).
[21] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[22] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási, Nature (London) **407**, 651 (2000).
[23] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103(R) (2003).
[24] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, *et al.*, Nature (London) **440**, 637 (2006).
[25] V. Colizza, A. Flammini, A. Maritan, and A. Vespignani, Physica A **352**, 1 (2005).
[26] Internet Network: undirected, unweighted network of the Internet at the Autonomous System level from data collected by the Oregon Route Views Project (http://www.routeviews.org/) in May 2001, where vertices represent Internet service providers and edges connections among them. The file reports the list of connected pairs of nodes.
[27] X. Guardiola, R. Guimerà, A. Arenas, A. Díaz-Guilera, D. Streib, and L. A. N. Amaral, e-print arXiv:cond-mat/0206240.
[28] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, Phys. Rev. E **70**, 056122 (2004).
[29] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, Behav. Res. Methods Instrum. Comput. **36**, 402 (2004).
[30] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[31] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[32] See EPAPS Document No. E-PLEEE8-78-103808 for supplementary material, which includes the in-depth characterisation of the computer-generated networks, the stability analysis of Eq. (2), the explanation for the degeneracy of l, details on the calculation of Q_{pred}, correlation coefficients of optimal values of l and topological properties of the network, and a representation of the metabolic network of *E. coli*.. For more information on EPAPS, see http://www.aip.org/pubserv/epaps.html.
[33] J. E. Hirsch, Proc. Natl. Acad. Sci. U.S.A. **102**, 16569 (2005).
[34] P. Ball, Nature (London) **448**, 737 (2007).
[35] A. Lancichinetti, S. Fortunato, and F. Radicchi, e-print arXiv:0805.4770.
[36] M. Kanehisa and S. Goto, Nucleic Acids Res. **28**, 27 (2000).
[37] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, Nucleic Acids Res. **30**, 42 (2002).

Supplementary Material for: The multistep greedy algorithm identifies community structure in real-world and computer-generated networks

Philipp Schuetz and Amedeo Caflisch[1]

[1]*Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland*

Fax: +41 44 635 68 62, Email: caflisch@bioc.uzh.ch

(Dated: July 4, 2008)

Contents

I. PROPERTIES OF COMPUTER-GENERATED NETWORKS

The diversity among the computer-generated networks is illustrated in Table I. To complement the main text (where only $l = 0.251\sqrt{L}$ is tested with L the number of edges), the modularity $Q_{\rm pred}$ results are displayed for the MSG-VM application with l as in Eq. (2) (six values are tested, i.e. $l_\alpha = \alpha\sqrt{L},\ \alpha = 0.25, 0.5, 0.75, 1$ and the two adjacent integers to l_α yieldings the highest Q value). The value $Q_{\rm pred}$ is smaller than the expectation value $\langle Q_{\rm rand}^{l<1.5\sqrt{L}}\rangle$ only for 32 (of 300), 19 (of 200), and 15 (of 300) networks of type *SED*, *SLD*, and *LLD*, respectively.

Network					MSG-VM with optimal l				MSG-VM with l from Eq. (2)			MSG-VM with random l	
Type	Realization Index	Vertices	Edges (L)	$\bar{k} \pm \sigma$	$l_{opt}/\sqrt{L}$	Q_{opt}	C_{opt}	$\bar{n} \pm \sigma$	Q_{pred}	C_{pred}	$\bar{n} \pm \sigma$	$\langle Q_{\text{rand}} \rangle$	$\langle Q_{\text{rand}}^{l<1.5\sqrt{L}} \rangle$
SED	1	264	284	2.2 ± 4.0	1.19	0.789	22	12 ± 8.5	0.789	22	12.0 ± 9.0	0.789	0.789
SED	2	467	486	2.1 ± 1.1	0.59	0.878	20	23.3 ± 7.6	0.878	20	23.3 ± 7.6	0.878	0.878
SED	3	346	793	4.6 ± 2.4	1.21	0.494	12	28.8 ± 8.4	0.488	11	31.4 ± 5.5	0.484	*0.490*
SED	4	322	817	5.1 ± 14.8	0.94	0.365	10	32.2 ± 18.3	0.364	9	35.7 ± 15.4	0.358	0.363
SED	5	550	942	3.4 ± 14.7	0.68	0.502	16	34.3 ± 29.5	0.502	16	34.3 ± 29.5	0.497	0.501
SED	6	301	1299	8.6 ± 11.5	0.31	0.290	8	37.6 ± 16.4	0.290	8	37.6 ± 16.4	0.284	0.285
SED	7	774	1579	4.1 ± 3.2	0.20	0.531	17	45.5 ± 8.9	0.527	17	45.5 ± 11.3	0.521	0.526
SED	8	636	1699	5.3 ± 17.4	0.19	0.388	12	53 ± 27.3	0.387	14	45.4 ± 23.7	0.379	0.385
SED	9	726	2208	6.1 ± 12.4	0.53	0.377	13	55.8 ± 27.3	0.377	13	55.8 ± 27.3	0.369	0.375
SED	10	513	2716	10.6 ± 23.7	1.30	0.232	8	64.1 ± 23.9	0.230	9	57.0 ± 17.1	0.228	0.230
SED	11	902	3191	7.1 ± 17.0	0.27	0.348	10	90.2 ± 38.9	0.348	10	90.2 ± 38.9	0.338	0.344
SED	12	657	3601	11.0 ± 11.9	0.32	0.270	9	73 ± 33.1	0.266	8	82.1 ± 33.8	0.261	0.266
SED	13	846	3984	9.4 ± 27.4	0.24	0.261	9	94 ± 40.6	0.261	9	94.0 ± 40.6	0.252	0.257
SED	14	743	4914	13.2 ± 10.9	0.66	0.249	10	74.3 ± 18.5	0.248	9	82.5 ± 25.4	0.241	0.245
SED	15	513	2716	10.6 ± 23.7	1.30	0.232	8	64.1 ± 23.9	0.230	9	57.0 ± 17.1	0.228	0.230
SLD	1	289	940	6.5 ± 7.2	0.23	0.374	9	32.1 ± 13.8	0.374	9	32.1 ± 13.8	0.364	0.368
SLD	2	995	1242	2.5 ± 3.0	0.14	0.768	26	38.2 ± 14.3	0.767	27	36.8 ± 13.1	0.764	0.766
SLD	3	1640	2465	3.0 ± 17.3	0.50	0.613	31	52.9 ± 49.7	0.613	31	52.9 ± 49.7	0.610	0.612
SLD	4	2211	3554	3.2 ± 41.7	0.37	0.445	25	88.4 ± 193	0.445	23	96.1 ± 200	0.443	0.444
SLD	5	878	3841	8.7 ± 5.0	0.26	0.324	9	97.5 ± 35.4	0.324	9	97.5 ± 35.4	0.314	0.320
SLD	6	1540	5226	6.8 ± 44.0	0.72	0.261	12	128.3 ± 88.4	0.260	11	140 ± 113.9	0.257	0.258
SLD	7	1117	5270	9.4 ± 23.0	0.34	0.285	9	124.1 ± 53.3	0.284	10	111.7 ± 69.8	0.276	0.282
SLD	8	485	5348	22.1 ± 31.2	0.23	0.152	5	97 ± 9.4	0.151	6	80.8 ± 26.4	0.145	0.148
SLD	9	1591	5432	6.8 ± 4.9	0.05	0.377	12	132.5 ± 67.9	0.375	14	113.6 ± 26.1	0.365	0.373
SLD	10	2242	7993	7.1 ± 47.3	0.07	0.284	13	172.4 ± 117	0.280	16	140.1 ± 118.8	0.279	0.280
SLD	11	904	11668	25.8 ± 25.0	0.08	0.166	6	150.6 ± 64.4	0.163	6	150.6 ± 34.3	0.156	0.161
SLD	12	721	11865	32.9 ± 33.0	0.39	0.136	7	103 ± 20.2	0.135	6	120.1 ± 52.8	0.129	0.133
SLD	13	1689	15722	18.6 ± 47.5	0.43	0.178	7	241.2 ± 60.1	0.176	6	281.5 ± 55.6	0.168	0.175
SLD	14	2992	25525	17.1 ± 50.4	0.08	0.196	8	374 ± 244.7	0.195	8	374 ± 220.7	0.185	0.193
SLD	15	3393	26257	15.5 ± 8.8	0.20	0.233	10	339.3 ± 246.2	0.231	10	339.3 ± 203.4	0.221	0.229
LLD	1	544	2084	7.7 ± 27.5	0.39	0.236	8	68 ± 29.5	0.235	8	68 ± 28.4	0.232	0.235
LLD	2	438	5061	23.1 ± 28.0	0.84	0.151	7	62.5 ± 11.1	0.151	6	73 ± 9.1	0.146	0.149
LLD	3	1469	6841	9.3 ± 44.0	0.08	0.232	10	146.9 ± 70.9	0.228	11	133.5 ± 66.5	0.226	0.227
LLD	4	594	6818	23.0 ± 27.6	0.33	0.167	7	84.8 ± 34	0.165	6	99 ± 13.7	0.156	0.161
LLD	5	3869	7931	4.1 ± 61.6	0.94	0.367	21	184.2 ± 305.3	0.367	21	184.2 ± 304.8	0.366	0.366
LLD	6	590	8086	27.4 ± 32.1	0.26	0.144	5	118 ± 9.2	0.144	5	118 ± 9.2	0.136	0.140
LLD	7	3554	8609	4.8 ± 58.6	3.21	0.345	22	161.5 ± 205.7	0.344	21	169.2 ± 222.2	0.342	0.343
LLD	8	2281	12298	10.8 ± 58.0	0.04	0.210	10	228.1 ± 97.7	0.207	13	175.4 ± 86.9	0.203	0.206
LLD	9	4193	19992	9.5 ± 66.5	0.11	0.249	12	349.4 ± 212.3	0.248	13	322.5 ± 200.6	0.235	0.244
LLD	10	1002	26622	53.1 ± 35.5	0.28	0.113	6	167 ± 56.1	0.110	6	167 ± 22.6	0.107	0.110
LLD	11	1133	36783	64.9 ± 41.2	0.34	0.101	6	188.8 ± 40.8	0.099	5	226.6 ± 40.5	0.096	0.098
LLD	12	3485	42297	24.3 ± 53.1	0.32	0.166	8	435.6 ± 188	0.164	7	497.8 ± 262.8	0.155	0.163
LLD	13	2335	54802	46.9 ± 55.5	0.34	0.116	7	333.5 ± 67.5	0.116	7	333.5 ± 53.4	0.108	0.114
LLD	14	3010	75776	50.3 ± 41.7	0.16	0.120	6	501.6 ± 174	0.119	7	430 ± 107.8	0.113	0.118
LLD	15	4165	127797	61.4 ± 50.6	0.26	0.108	5	833 ± 113.6	0.107	6	694.1 ± 323.2	0.101	0.106

TABLE I: Heterogeneity of computer-generated networks and comparison of MSG-VM results using l as in Eq. (2) of main text or picked at random. For each of the three network types, 15 realizations are shown ranked by L. The degree heterogeneity is evident in the average and standard deviation of the degree (column "$\bar{k} \pm \sigma$"). The column "Q_{opt}" lists the maximal value of modularity obtained by running MSG-VM for all values of l smaller than $\min\{5000, L\}$ (L the number of edges). The column "Q_{pred}" lists the MSG-VM modularity obtained using Eq. (2) of the main text to determine the step width. The columns "C_{opt}" and "C_{pred}" list the number of communities in the partitions with modularity Q_{opt} and Q_{pred}, respectively. The average and standard deviation of the number of vertices per community are listed in the columns "$\bar{n} \pm \sigma$". The columns "$\langle Q_{\text{rand}} \rangle$" and "$\langle Q_{\text{rand}}^{l<1.5\sqrt{L}} \rangle$" show the expectation value for the MSG-VM modularity when six values of l are picked randomly from a uniform distribution in the range $1 \leq l \leq \min\{5000, L\}$ and $1 \leq l \leq 1.5\sqrt{L}$, respectively. The expectation value is estimated by averaging, over 1000 samples, the highest modularity obtained using six values of l (confer section VII for details). A total of six values of l are picked randomly because six values were used to determine Q_{pred}: the four values of l calculated by Eq. (2) of the main text and the two integers adjacent to the best of these four. There is only one value of $\langle Q_{\text{rand}} \rangle$ and $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}} \rangle$ higher than the corresponding Q_{pred} (in italics).

Label	#	Deg.	$\frac{l_{\text{opt}}}{\sqrt{L}} > 1.5$	VM-effect	MSG-effect
GN_1	100	61	13 %	13	0
GN_2	100	35	15 %	12	3
GN_3	100	71	18 %	18	0
SED	300	54	7.3 %	19	3
SLD	200	23	5.5 %	9	2
LLD	300	10	0.7 %	0	2

TABLE II: Statistical properties of l_{opt} (smallest value of step width that yields the highest MSG-VM modularity) for the computer-generated networks. The column *Deg.* lists the number of networks for which multiple values of l yield $Q_{\text{MSG-VM}}\,(l_{\text{opt}})$. The fraction of examples with $l_{\text{opt}} > 1.5\sqrt{L}$ (L total edge weight) is small. These networks are classified according to the *VM-labels* (confer Sec. III A for details): "VM-effect" and "MSG-effect".

II. SOURCES OF DEGENERACY

For multiple computer-generated networks more than one value of step width yield the highest MSG-VM modularity (Table II). In contrast, all real-world networks with three exceptions (the jazz, the metabolic *E. coli*, and the Zachary karate club network) have a unique optimal value l_{opt} of the step width. For the Girvan-Newman networks $GN_{1,2,3}$ the number of l_{opt} values displays a phase-transition like behavior (Fig. 1) upon variation of z_{out} (average number of edges connecting a vertex in one of the four imposed communities to members of another module). The transition between the two "phases" (low z_{out} value with many l_{opt} values and high z_{out} value with few l_{opt}) occurs for similar values of $\frac{z_{\text{out}}}{\text{max. degree}}$ whereas the fraction becomes larger with increasing network size. In networks with small z_{out} value many vertices and their neighborhoods are similar in contrast to graphs with high z_{out} value. Therefore, this symmetry (almost identity of vertices) is assumed to be the source of degeneracy.

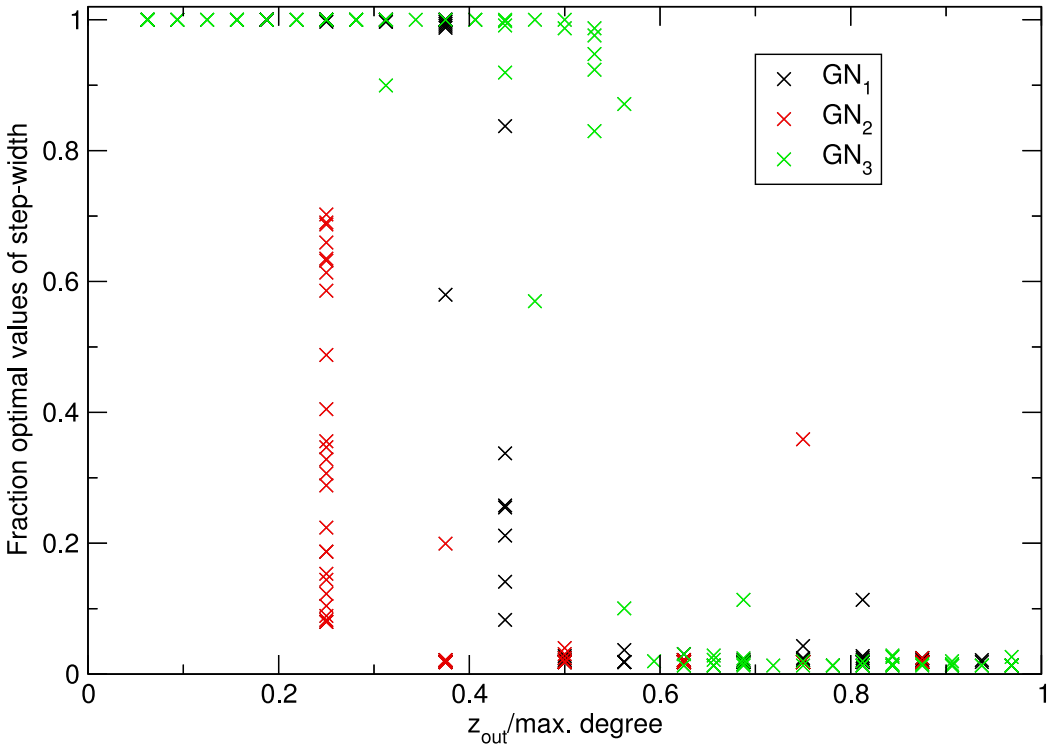

FIG. 1: Influence of z_{out} (average number of inter-community edges per vertex) on the number of distinct values of step width yielding the highest MSG-VM modularity in the Girvan-Newman networks $GN_{1,2,3}$.

III. RESTRICTABILITY OF SEARCH RANGE

A. Quantification of VM-Contribution

The best MSG-VM solution can emerge from two disparate scenarios: First, an excellent MSG solution is insignificantly finetuned by the VM procedure (*MSG-effect*). Second, the VM bears the lion's share and optimizes a poor MSG solution (*VM-effect*). To discern these two cases, the following criteria are defined (l_{opt} is smallest value of the step width that maximizes the MSG-VM modularity, $Q_{\text{MSG}}(l)$ the modularity after the application of the MSG algorithm with l as step width):

"MSG-effect" The modularity obtained by the MSG algorithm with l_{opt} as step width is at least 90 % of the maximal MSG modularity (among all other tested values of step width):

$$Q_{\text{MSG}}(l_{\text{opt}}) > 0.9 \max_{l}(Q_{\text{MSG}}(l))$$

and among the best 50 % of all Q_{MSG}-values:

$$Q_{\mathrm{MSG}}(l_{\mathrm{opt}}) > \frac{\min_l \left(Q_{\mathrm{MSG}}(l)\right) + \max_l \left(Q_{\mathrm{MSG}}(l)\right)}{2}.$$

The second condition assures that no fallacious MSG-dominance is identified when the MSG modularity values fluctuate less than 20 % upon variation of the step width.

"VM-effect" All other cases.

B. Optimal parameter smaller than $1.5\sqrt{L}$

If the MSG algorithm dominates the optimization, the arguments in section III.A of the main text imply that the optimal value of the step width l_{opt} (smallest integer that yields the highest MSG-VM modularity) should be smaller than $\beta\sqrt{L}$ (L total edge weight and β a prefactor on the order of 1). For the VM-driven examples no such concentration can be expected. To test this hypothesis, the computer-generated networks are splitted according to the criteria in section III A. For both groups, a histogram of $r = \frac{l_{\mathrm{opt}}}{\sqrt{L}}$ is calculated (not shown). In agreement with the hypothesis the MSG distribution drops significantly between $r = 1$ and $r = 2$ independent of the network type. Choosing $r_{\mathrm{th}} = 1.5$ as threshold, 92.6 % of the computer-generated networks have $l_{\mathrm{opt}} < r_{\mathrm{th}}\sqrt{L}$. Among these 1019 networks only 41 are VM-driven. 71 of 81 networks with $l_{\mathrm{opt}} > r_{\mathrm{th}}\sqrt{L}$ are VM-driven (Table II). The cutoff value $r_{\mathrm{th}} = 1.5$ demarcates the boundary between MSG- and VM-driven networks. Furthermore, the optimal value of the step width is expected to be an integer smaller than $1.5\sqrt{L}$.

IV. EFFECTS OF VERTEX LABELING PERMUTATION

Permuting the vertex labels leaves the topology invariant, but changes the order in which the MSG-VM algorithm parses the vertices. Thus, the influence of non-topological contributions on the optimization can be studied. Furthermore, by averaging over the profiles of the same network with different vertex labellings the intrinsic accuracy limit of a prediction procedure based on topological properties can be determined. Here, 100 variants of the smallest 10 real-world networks with different vertex labellings are created. On each copy the MSG-VM algorithm is applied for all possible values of step width (i.e. all integers

				VM-Label	
Network	# l_{opt}	>	Δ[%]	VM	MSG
Zachary	78	0	0.00	0	100
Metabolic *E. coli*	49	24	0.12	3	97
College Football	400 [a]	12	0.18	17	83
Metabolic *C. elegans*	141	50	0.91	25	75
Jazz	700 [b]	0	0.00	79	21 [c]
Email	94	80	0.57	15	85
Yeast (PPI, LCC)	18	4	0.05	0	100
M. Karplus	107	36	0.94	4	96
PPI *S. cerevisiae* (LCC)	56	69	0.46	0	100
PPI *S. cerevisiae*	56	82	0.52	0	100

[a] Omitting the networks with "VM-effect" label only 20 different values are found.
[b] If the networks with "VM-effect" are omitted, only six distinct values are found.
[c] Only 15 networks have exclusively l_{opt} values with "MSG-effect" label.

TABLE III: Effect of permutation of vertex labels: For the smallest 10 real-world networks 100 copies with scrambled vertex labels are created. On each copy the MSG-VM procedure is applied using all integers smaller than the number of edges as step width. The column $\#l_{opt}$ lists the number of distinct values of the step width yielding the highest MSG-VM modularity. The number of copies for which the highest MSG-VM modularity is higher than the one for the unscrambled variant is listed in column ">". The highest improvement is shown in column "Δ". The columns "VM-label" report on the number of copies for which the optimization is "MSG" or "VM" dominated.

smaller than the number of edges). The maximal MSG-VM modularity improves at most by 0.94% in comparison to the unscrambled variant (Table III). This change is marginal and smaller than the modularity change upon variation of the step width. Therefore, an empirical formula using topological properties to predict l_{opt} has an accuracy limit of at least one percent.

Strikingly, whether the VM or the MSG dominates the optimization is conserved under vertex label permutation. The excessive diversity of the optimal values of l for the *college football* and the *jazz* network originates from the VM-driven optimization. By taking the

Label	Mean [%]	$\alpha_{\max}$	Median [%]	α_{median}
Zachary Karate Club	100.00	0.57	100.00	0.57
Metabolic *E. coli*	99.90	0.047	99.92	0.47
College Football	98.85	0.080	99.82	0.040
Metabolic *C. elegans*	98.91	0.71	99.00	0.71
Jazz	99.96	0.57	99.97	0.52
Email	99.51	0.65	99.56	0.65
Yeast (PPI, LCC)	99.37	0.27	99.43	0.27
M. Karplus	99.10	0.72	99.17	0.72
PPI *S. cerevisiae* (LCC)	99.55	0.48	99.60	0.48
PPI *S. cerevisiae*	99.49	0.49	99.56	0.49

TABLE IV: Effect of permuted vertex labels on location and value of the peak in the MSG-VM modularity curve. The average/median of the MSG-VM modularity profiles $\bar{Q}_{\text{MSG-VM}}$ is calculated over 100 copies of the smallest ten real-world networks networks with permuted vertex labels. For the mean and median curve the peaks are located at the relative parameter $\alpha_{\max}$ and α_{median}, respectively.

mean or the average over the $\bar{Q}_{\text{MSG-VM}}$ (α) profiles (defined in Sec. III.A.1. of the main text) of the 100 variants with scrambled vertex labels the non-topological contributions are smoothed out. Remarkably, the resulting average/median profiles peak for almost all networks (two exceptions with $l_{\text{opt}} = 1$ and the email network) at values of α in close vicinity of multiples of 0.25 (Table IV).

V. STABILITY OF SCALING FACTOR

The MSG-VM algorithm performs best (on average over all computer-generated networks), if $\lfloor \alpha\sqrt{L} \rfloor$ with $\alpha = 0.251$ is chosen as value of the step width. To assess the effect of another selection of the network set, a leave-23-out test is performed. For this test 10000 samples of 980 out of 1003 computer-generated networks (the 97 networks for which the MSG-VM modularity is independent of the chosen step width are excluded) and for each sample $\bar{Q}_{\text{MSG-VM}}(\alpha)$ is calculated. In Fig. 2 the average and standard deviation

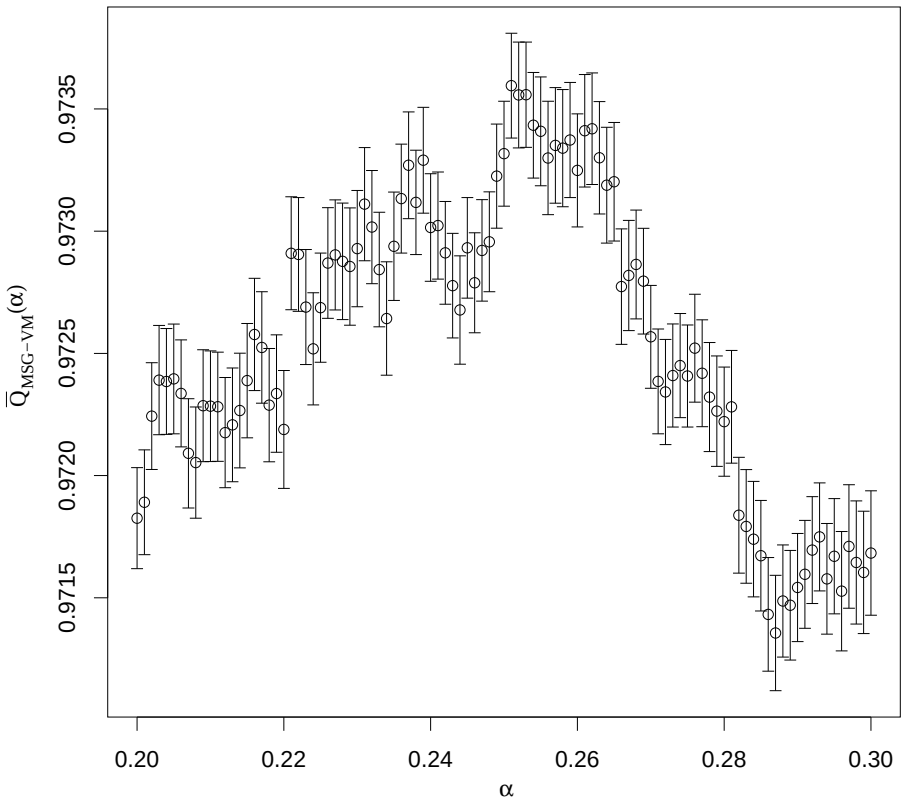

FIG. 2: Leave-23-out procedure on $\bar{Q}_{\text{MSG-VM}}(\alpha)$ curve (definition in Sec. III.A.1 of main text): 10000 samples of 980 out of 1003 computer-generated networks (97 network are excluded as they display no dependence on different values of the step width) are taken and the statistics over the corresponding $\bar{Q}_{\text{MSG-VM}}(\alpha)$ curves is displayed. The peak is at $\alpha = 0.251$. Furthermore, for $\alpha = 0.252, 0.253$ the one-σ-ranges overlap considerably.

of all $\bar{Q}_{\text{MSG-VM}}$ curves are shown. Remarkably, the average $\bar{Q}_{\text{MSG-VM}}$ (average over 10000 samples) peaks at $\alpha = 0.251$, too. At $\alpha = 0.252, 0.253$ the average $\bar{Q}_{\text{MSG-VM}}(\alpha)$ values are within the standard deviation of the values at $\alpha = 0.251$. Reminding that $\alpha = 0.251, 0.252, 0.253$ predict basically the same value of the step width for networks with less that 10^6 edges, this variance in α is negligible.

VI. DETAILS OF Q_{pred} CALCULATION

The modularity Q_{pred} in Table I of the main text is calculated as

$$Q_{\text{pred}} = Q_{\text{MSG-VM}}\left(\left\lfloor \alpha\sqrt{L} \right\rfloor + l_\alpha\right)$$

Network	α	Δl_α
Zachary Karate Club	0.25, 0.5, 0.75	+1
Metabolic *E. coli*	0.25	0
College Football	0.75, 1	−1
Metabolic *C. elegans*	0.75	−1
Jazz	0.75, 1	0
Email	0.75	+1
Yeast (PPI, LCC)	1	0
M. Karplus	0.75	0
PPI *S. cerevisiae* (LCC)	0.5	+1
PPI *S. cerevisiae*	0.5	−1
Internet	0.5	−1
PGP-key signing	0.25	+1
Word Association (LCC)	1	−1
Word Association	0.75	+1
Collaboration	0.5	−1
WWW	0.75	0
Actor	0.75	+1

TABLE V: Parameters for $Q_{\text{pred}} = Q_{\text{MSG-VM}}\left(\left\lfloor \alpha\sqrt{L}\right\rfloor + l_\alpha\right)$ calculation (L the number of edges) used in Table I of the main text.

with α,l_α as given in Table V in the supplementary material and L the number of edges.

VII. CALCULATION OF $\langle Q_{\text{rand}}\rangle$ AND $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}}\rangle$

The expectation values for a random selection of the step width (Table I of the main text) $\langle Q_{\text{rand}}\rangle$ and $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}}\rangle$ are not calculated accurately. These values are approximated by the averages over 1000 sampling experiments in which the highest modularity is calculated for six values of the step width picked at random. The six values of the step width are selected from all tested values for $\langle Q_{\text{rand}}\rangle$ and from those values of step width smaller than $1.5\sqrt{L}$ (L total edge weight) for $\langle Q_{\text{rand}}^{l<1.5\sqrt{L}}\rangle$.

VIII. CORRELATIONS OF l_{opt} WITH TOPOLOGICAL PROPERTIES

Table VI displays a selection of correlation values of topological properties and powers thereof with l_{opt} (the smallest value of the step width yielding the highest MSG-VM modularity). In this calculation only the 905 computer-generated network with a unique l_{opt} are included to reduce the effect of ambiguities. The powers $0.1, 0.2, \ldots, 2$ are considered. Higher powers are excluded to reduce the danger of overfitting. Irrespective of the power considered, the quantities "vertices" and "edges" correlate much better with l_{opt} than all other properties.

IX. EXHAUSTIVE COMPARISON OF MSG-VM AND GREEDY PARTITION FOR METABOLIC NETWORK OF *E. COLI*

The differing parts between the MSG-VM and the greedy partition of the metabolic *E. coli* network are shown in Fig. 3 and 4 (identically classified vertices and edges to/among them are removed). The vertices of the pathway "Puridine Metabolism" (green shaded area) are put in one community by both algorithms. However, the greedy algorithm merges these vertices with those around vertex "C00049" (Aspartate) whereas the MSG-VM algorithm separates them. As aspartate is not involved in the "Puridine Metabolisme", this merge performed by the greedy algorithm is artificial. The MSG-VM algorithm gathers almost all metabolites (the exceptional vertex "C00084" stands for "Acetaldehyde") belonging to the "Glycerophospholipide Metabolism" (yellow shaded area) in one community. In contrast, the greedy algorithm separates the vertices "C01233" (Glycerophosphoethanolamine) and "C000093" (Glycerol-3-phosphate) in addition. Acetaldehyde is part of multiple pathways and therefore might be assigned to other pathways as well. The metabolite "C01233" belongs to only one pathway and therefore is misplaced by the greedy algorithm. Glycerol-3-phosphate belongs to the "Glycerolipid metabolism" pathway to which the vertex "C00577" is assigned as well. To summarize, the greedy algorithm merges artificially two modules and misassignes three vertices of which one is uniquely assigned. The MSG-VM approach misplaces for these two examples only one vertex.

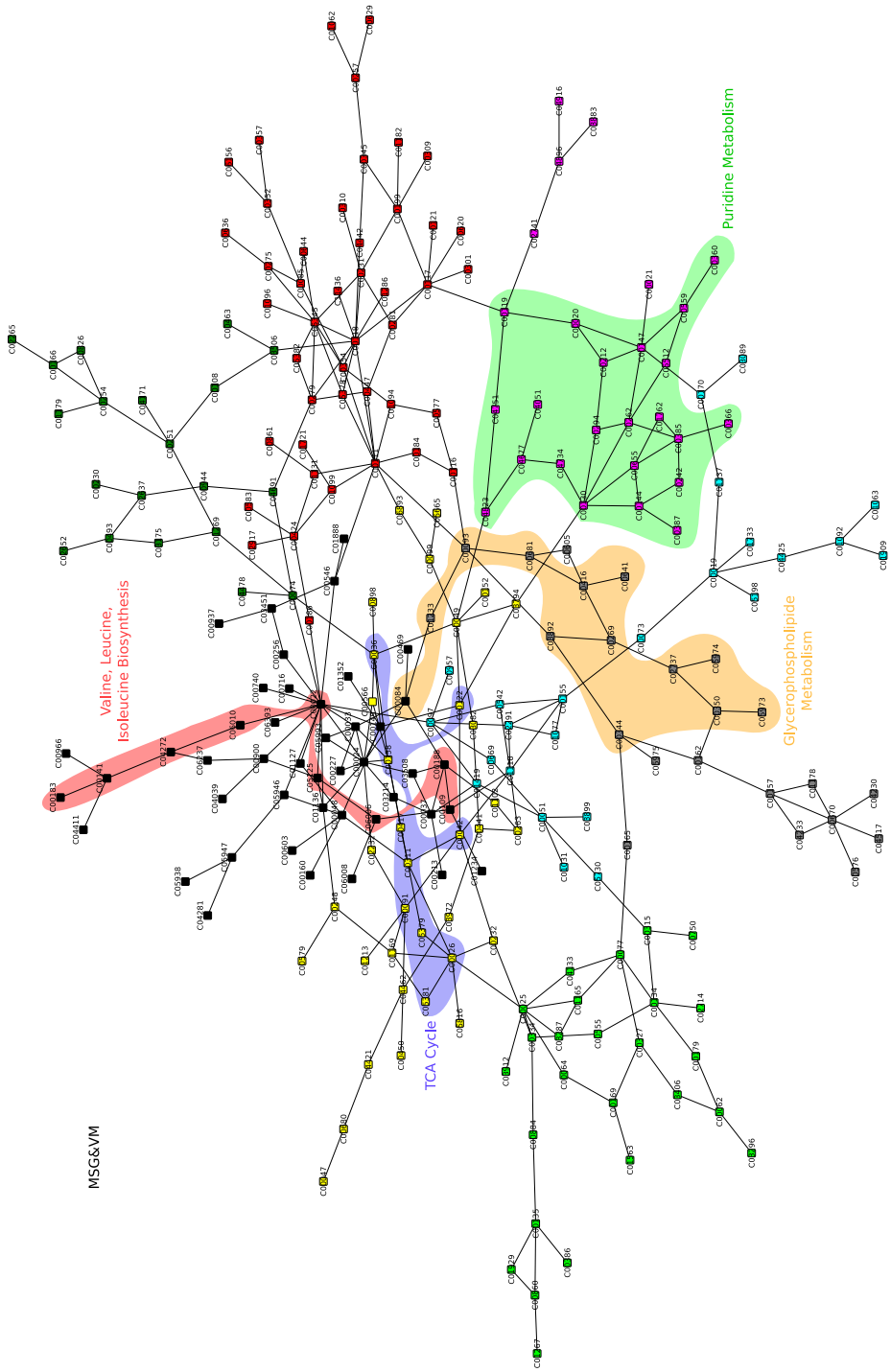

FIG. 3: Partition of the metabolic network of *E. coli* obtained by the MSG-VM algorithm with $l = 6$. The network is reduced on those vertices and edges belonging to communities that differ between the MSG-VM and the greedy solution. The vertex labels are taken from the KEGG database. For instance, detailed information about vertex C00149 can be found at `http://www.genome.jp/dbget-bin/www_bget?compound+C00149`).

	Power			
	0.5	1	1.5	2
Vertices	0.6498	0.6599	0.6482	0.6289
Edges	0.7728	0.7314	0.6669	0.6067
$\langle$Degree$\rangle$	0.6584	0.6596	0.6350	0.6009
Max. Degree	0.4419	0.3589	0.2789	0.2153
σ(Degree)	0.6428	0.6538	0.6384	0.6123
$\langle$CC$\rangle$	-0.0339	-0.0424	-0.0390	-0.0325
σ(CC)	-0.3363	-0.2877	-0.2335	-0.1838
$\langle$CC$^2\rangle$	-0.1021	-0.0583	-0.0362	-0.0266
σ(CC^2)	-0.1944	-0.1876	-0.1647	-0.1366
$\langle$CC$^3\rangle$	-0.1225	-0.0573	-0.0355	-0.0278
σ(CC^3)	-0.1266	-0.1587	-0.1527	-0.1296
$\langle\Delta$CC$\rangle$	-0.3013	-0.2024	-0.1318	-0.0915
$\sigma(\Delta$CC)	-0.4326	-0.3260	-0.2340	-0.1660
$\langle\Delta$Degree$\rangle$	0.2719	0.1318	0.0580	0.0352
$\sigma(\Delta$Degree)	0.2294	0.1117	0.0532	0.0297

TABLE VI: Correlations of topological properties and powers thereof with l_{opt} (value of step width that yields the highest MSG-VM modularity). $\langle x\rangle$ indicates that the average of property x over all vertices is considered. The standard deviation of property y over all vertices is abbreviated by $\sigma(y)$. "CC" stands for clustering coefficient. $\langle\Delta$Degree$\rangle$ and $\langle\Delta$CC$\rangle$ denotes the average over the differences in degree and clustering coefficient of two linked vertices, respectively.

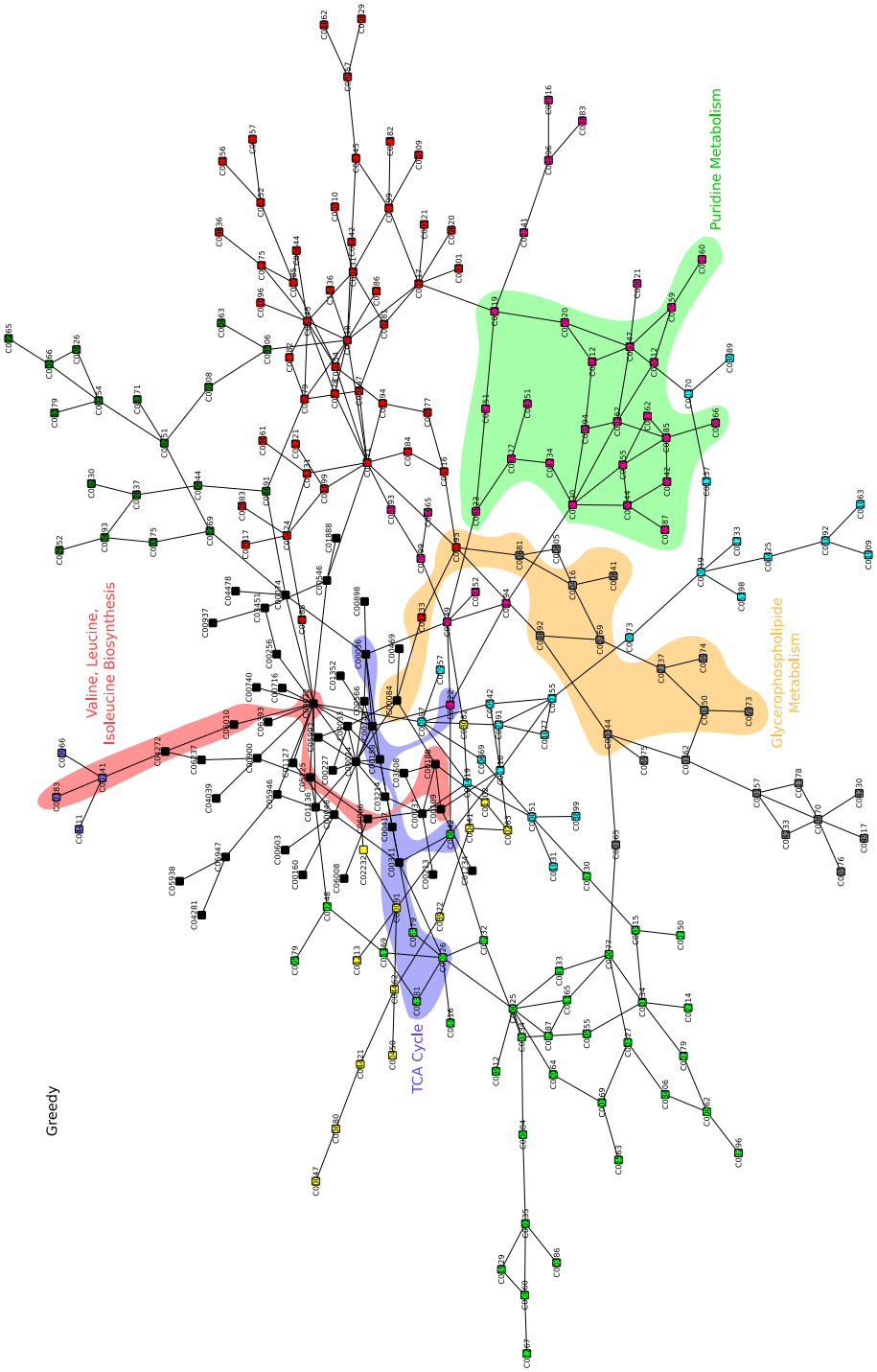

FIG. 4: Clusterization of the metabolic network of *E. coli* obtained by the greedy algorithm. The same excerpt as in Fig. 3 is displayed.

Conclusions and Outlook

Molecular dynamics (MD) simulations model protein dynamics at full atomistic level of detail. Recent methods characterize the free-energy landscape from MD data using networks as underlying technology [1–5]. Experimentally, "Förster Resonance Energy Transfer" (FRET) can monitor a single intramolecular distance with up to nanosecond time resolution over an extended period of time [6, 7]. A key question addressed in this thesis is how much information on the free-energy surface can be gained from a time-series of an one-dimensional observable, e.g. an intramolecular distance.

The first part of this thesis introduces FESST (Free Energy Surface from Single-molecule Time series), a procedure to infer free-energy basins from single-molecule data. The key idea is that the distribution of the signal in a short time window is characteristic for the current state of the system. The free-energy basins are then extracted from the trajectory of system states by cut-based free-energy profiles (cFEPs) [3, 5]. Applied to Beta3s [8], FESST extracts multiple free-energy basins and the height of the barrier between folded and unfolded state with high accuracy from the time series of a single distance (cf. chapter 2). For a simulated FRET experiment, FESST extracts a mainly native basin from the time series of FRET efficiencies for as few as 200-1000 photons per folding time. Extrapolated to single molecule experiments, distinct free-energy basins can be detected for proteins with a folding time of few milliseconds. Although the high threshold prohibits the analysis of fast folders [9], many proteins display dynamics on the time scale accessible to FESST [10–13]. Therefore, future research needs to address the following three issues: Reduction of the required amount of photons, elimination of experimental noise, and combination of information from multiple short bursts. The threshold on the relaxation time can be lowered if another observable is identified that requires fewer photons for its definition. A candidate for this type of observable is the transfer efficiency calculated

from inter-photon times. Experimental noise may be reduced by a careful selection of the photon bursts considered for the analysis and a smoothing procedure for the time series of the observable. For instance, wavelets removed successfully the effect of thermal noise in the inferral of a state space network from an one-dimensional signal [14]. Due to photophysical effects (bleaching of chromophores) and/or diffusion of the proteins, the recorded photon traces are shorter than the trajectories extracted from MD simulations. To obtain an encompassing picture of the protein dynamics, the traces of multiple molecules have to be combined, for instance by a statistical procedures as suggested by Li et al. [14] and Komalapriya et al. [15]. Finally, it is important to notice that the key idea of FESST to compare the short-time behavior of an observable can be used more generically to distinct states in noisy low-dimensional time series.

The second part of this thesis focuses on the development and improvement of procedures to identify densely connected groups of nodes in large networks. The devised procedure MSG-VM (multistep-greedy and vertex-mover algorithm) optimizes modularity, an objective assessment function, on a set of established benchmark networks better and with smaller running time than previous procedures [16, 17]. The pathways detected by MSG-VM in the metabolic network of *Escherichia coli* are more consistent with the literature annotation than those detected by the greedy algorithm [18]. As a caveat, it has been shown that modularity is not able to resolve a community smaller than a given size [19]. All recently published procedures [18, 20, 21] rely on the stepwise rearrangement of few nodes or the local optimization of a modified cost function. For future research, a promising strategy might be to combine well-known building blocks of communities to obtain high qualitative partitions in short running time. The development of clusterization techniques allows the treatment of networks with increasing size (several million vertices and edges are feasible even today). Noteworthily, a clusterization of a network not only reveals hidden substructures, but can also be used to examine changes of the network in time (a modified community structure is caused by a change in the local topology).

In summary, the projection of data on networks enables the analysis of data from disparate fields with the same formalism. Furthermore, the common framework allows an immediate transfer of tools developed for one analysis to other fields of application.

Bibliography

[1] S. V. Krivov and M. Karplus. Free energy disconnectivity graphs: Application to peptide models. *J. Chem. Phys.*, 117:10894–10903, 2002.

[2] S. V. Krivov and M. Karplus. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.*, 101:14766–14770, 2004.

[3] S. V. Krivov and M. Karplus. One-dimensional free-energy profiles of complex systems: progress variables that preserve the barriers. *J. Phys. Chem. B*, 110:12689–12698, 2006.

[4] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins*, 70:1185–1195, 2008.

[5] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus. One-dimensional barrier-preserving free-energy projections of a beta-sheet miniprotein: new insights into the folding process. *J. Phys. Chem. B*, 112:8701–8714, 2008.

[6] A. J. Berglund, A. C. Doherty, and H. Mabuchi. Photon statistics and dynamics of fluorescence resonance energy transfer. *Phys. Rev. Lett.*, 89:068101, 2002.

[7] B. Schuler, E. A. Lipman, and W. A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature (London)*, 419:743–747, 2002.

[8] E. de Alba, J. Santoro, M. Rico, and M. A. Jimnez. De novo design of a monomeric three-stranded antiparallel beta-sheet. *Protein Sci.*, 8:854–865, 1999.

[9] A. Hoffmann, A. Kane, D. Nettels, D. Hertzog, P. Baumgärtel, J. Lengefeld, G. Reichardt, D. Horsley, R. Seckler, O. Bakajin, and B. Schuler. Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.*, 104:105–110, 2007.

[10] C. Joo, S. A. McKinney, M. Nakamura, I. Rasnik, S. Myong, and T. Ha. Real-time observation of reca filament dynamics with single monomer resolution. *Cell*, 126:515–527, 2006.

[11] A. Borgia, P. M. Williams, and J. Clarke. Single-molecule studies of protein folding. *Annual Review of Biochemistry*, 77(1):101–125, 2008.

[12] C. Joo, H. Balci, Y. Ishitsuka, C. Buranachai, and T. Ha. Advances in single-molecule fluorescence methods for molecular biology. *Annual Review of Biochemistry*, 77(1):51–76, 2008.

[13] H. S. Chung, J. Louis, and W. A. Eaton. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci. U.S.A.*, 106(29):11837–11844, 2009.

[14] C. Li, H. Yang, and T. Komatsuzaki. Multiscale complex network of protein conformational fluctuations in single-molecule time series. *Proc. Natl. Acad. Sci. U.S.A.*, 105:536–541, 2008.

[15] C. Komalapriya, M. Thiel, M. C. Romano, N. Marwan, U. Schwarz, and J. Kurths. Reconstruction of a system's dynamics from short trajectories. *Phys. Rev. E*, 78:066217, 2008.

[16] P. Schuetz and A. Caflisch. Multistep greedy algorithm identifies community structure in real-world and computer-generated networks. *Phys. Rev. E*, 78:026112, 2008.

[17] P. Schuetz and A. Caflisch. Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement. *Phys. Rev. E*, 77:046112, 2008.

[18] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, 2004.

[19] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proc. Natl. Acad. Sci. U.S.A.*, 104:36–41, 2007.

[20] R. Guimerà and L. A. N.Amaral. Functional cartography of complex metabolic networks. *Nature (London)*, 433:895–900, 2005.

[21] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72:027104, 2005.

Acknowledgments

This work would never have been possible without the help of many people. First of all, I would like to thank Prof. Amedeo Caflisch for his continuous support and the possibility to work on so many different aspects of biochemistry during my PhD studies. I would like to express my deepest gratitude to Prof. Ben Schuler for the numerous enlightening discussions and fruitful new perspectives.

All members of the Caflisch's and the Schuler's group helped to create a very friendly and helpful environment, in which it is a pleasure to work. Apart from work, many persons from music bands and other social activities helped me to find a good balance of science and other things. A special thanks goes to Rainer Böckmann, Christian Bolliger, Fabian Dey, Alexandros Guekos, Hagen Hofmann, Gianluca Interlandi, Peter Kolb, François Marchand, Steffi Muff and Daniel Nettels for innumerable discussion about science and other aspects of life.

Last, but not least I would like to thank my parents, Judith and Walter Schütz, and Susanne Wampfler for continuous support without which this thesis would not have been possible.

Curriculum Vitae

Philipp Schütz

born 12^{th} December 1978 in Aarau, Switzerland
Swiss citizen

Education

September 2004–present	PhD student at the Department of Biochemistry, University of Zurich.
April 2003 - September 2004	Postdiploma research at the Swiss Federal Institute of Technology in Zurich.
April 2003	Diploma in Physics at the Swiss Federal Institute of Technology in Zurich.
1998–2003	Studies of Physics at the Swiss Federal Institute of Technology in Zurich.
1993–1998	Kantonsschule Olten, grammar school specialising in ancient languages.